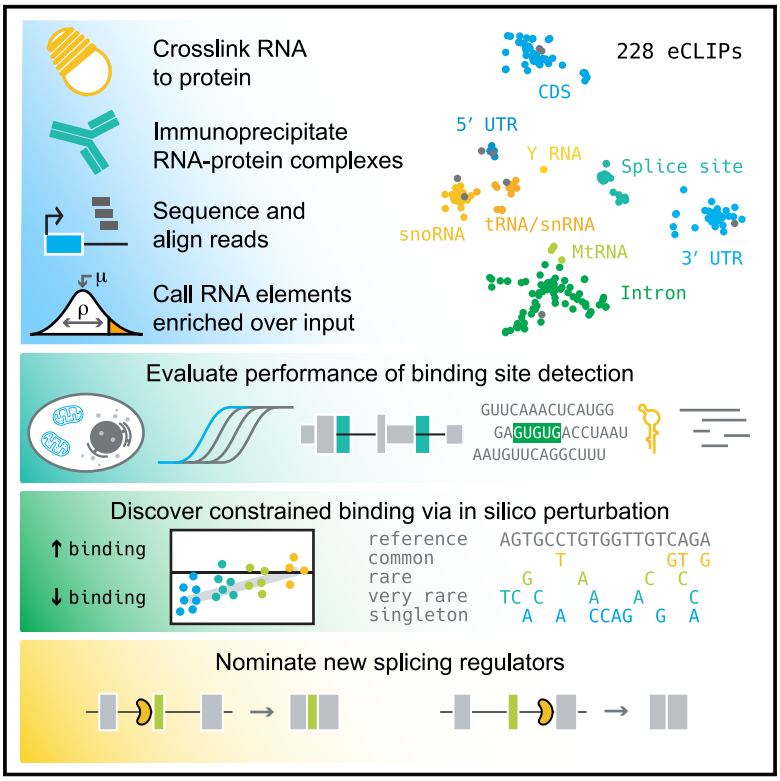
Cell Genomics

Technology

Skipper analysis of eCLIP datasets enables sensitive detection of constrained translation factor binding sites

Graphical abstract



Highlights

- Beta-binomial modeling improves detection of RNA-protein interactions from eCLIP data
- Reported binding sites include repetitive RNAs for 99% of studied proteins
- Ribosomal proteins and EIFs bind diverse RNA elements and sequence motifs
- Diamond-Blackfan anemia mutations alter RPS19's preferred transcriptomic targets

Authors

Evan A. Boyle, Hsuan-Lin Her, Jasmine R. Mueller, Jack T. Naritomi, Grady G. Nguyen, Gene W. Yeo

Correspondence

geneyeo@ucsd.edu

In brief

Boyle et al. introduce a framework for calling RNA-protein interactions transcriptome-wide from crosslinking and immunoprecipitation (CLIP) data. Their statistical approach quantifies binding to repetitive RNAs, detects >200% more total interactions, and attains greater accuracy than other tools. The resulting resource broadens the applicability of CLIP data for genome biology discovery.

Boyle et al., 2023, Cell Genomics *3*, 100317
June 14, 2023
https://doi.org/10.1016/j.xgen.2023.100317

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Technology

Skipper analysis of eCLIP datasets enables sensitive detection of constrained translation factor binding sites

Evan A. Boyle,[1] Hsuan-Lin Her,[1] Jasmine R. Mueller,[1] Jack T. Naritomi,[1] Grady G. Nguyen,[1] and Gene W. Yeo[1,2,*]
[1]Department of Cellular and Molecular Medicine, Institute for Genomic Medicine, UCSD Stem Cell Program, University of California San Diego, La Jolla, CA 92093, USA
[2]Lead contact
*Correspondence: geneyeo@ucsd.edu
https://doi.org/10.1016/j.xgen.2023.100317

SUMMARY

Technology for crosslinking and immunoprecipitation (CLIP) followed by sequencing (CLIP-seq) has identified the transcriptomic targets of hundreds of RNA-binding proteins in cells. To increase the power of existing and future CLIP-seq datasets, we introduce Skipper, an end-to-end workflow that converts unprocessed reads into annotated binding sites using an improved statistical framework. Compared with existing methods, Skipper on average calls 210%–320% more transcriptomic binding sites and sometimes >1,000% more sites, providing deeper insight into post-transcriptional gene regulation. Skipper also calls binding to annotated repetitive elements and identifies bound elements for 99% of enhanced CLIP experiments. We perform nine translation factor enhanced CLIPs and apply Skipper to learn determinants of translation factor occupancy, including transcript region, sequence, and subcellular localization. Furthermore, we observe depletion of genetic variation in occupied sites and nominate transcripts subject to selective constraint because of translation factor occupancy. Skipper offers fast, easy, customizable, and state-of-the-art analysis of CLIP-seq data.

INTRODUCTION

RNA-binding proteins (RBPs) conduct a vast array of essential functions in living cells. RNA synthesis, processing, modification, translation, and decay all require diverse RBPs that act in specific temporal, spatial, and cell type contexts.[1,2] Currently, crosslinking and immunoprecipitation (CLIP) followed by sequencing (CLIP-seq) methods are the gold standard for probing transcriptome-wide RNA-protein interactions in cells. However, CLIP-seq methods have continued to evolve and diversify,[3,4] requiring concomitant development of new tools for statistical modeling and data visualization.

CLIP-seq analysis must confront challenges inherent to analysis of both RNA sequencing (RNA-seq), where target sequences from transcript isoforms vary in expression by orders of magnitude, and chromatin immunoprecipitation sequencing (ChIP-seq), where read signal aggregates into peaks against roughly even unbound chromatin background. CLIP-seq analysis tools generally identify candidate binding sites by calling peaks[5–7] or modeling positional enrichment.[8–10] Regardless of the approach taken, few tools attempt to test for binding to multi-mapping sequences and repetitive elements in the human transcriptome,[7,11] even though repetitive elements are in some cases the principal targets of RBPs.[12]

Peak calling approaches are ill-suited to handle diverse RBP binding modes. Results from peak calling approaches are susceptible to false negatives due to masking of intronic signal by exonic reads at exon-intron boundaries,[13] bias against calling peaks on low-abundance transcripts,[10] and a mismatch between the length of bound regions and the bandwidth used for peak calling.[9] Apparent enrichment can vary either incrementally or abruptly over the course of a few nucleotides, but the sliding windows used to detect peaks are fixed in size. Even when binding profiles abide by expectations, reconciling partially overlapping peaks across samples and interpreting peaks that overlap multiple known transcripts is nontrivial.

Conversely, positional models can be underdetermined when binding sites are densely packed together and signals from distinct binding events overlap. More broadly, they require extensive parameterization that may be susceptible to biases related to CLIP-seq library fragment length or GC content. CLIP signal surrounding corresponding motifs often spans tens of nucleotides before decaying due to idiosyncratic patterns of RBP crosslinking to target sequences.[5,8] Cooperative and competitive RBP binding at nearby or overlapping binding sites can also alter RNA occupancy and regulatory outcomes.[14–16] Thus, interpretation of enrichment scores or binding affinity at nucleotide-level resolution remains challenging.

Here, we introduce an end-to-end solution for analyzing CLIP-seq data (Skipper) that skips peak calling by tiling windows over annotated transcripts. Skipper processes both uniquely mapping and multi-mapping reads to report bound elements

transcriptome-wide. Tiled windows and repetitive elements are tested for enrichment in immunoprecipitated over input samples using a beta-binomial distribution that accounts for overdispersion in read counts. We develop benchmarks for evaluating CLIP-seq data and compare Skipper with existing methods using enhanced CLIP (eCLIP) data available through the ENCODE project website. Furthermore, we demonstrate the broad applicability of Skipper output by collecting new eCLIP data on a medley of translation factors and identifying selective constraint acting on translation factor occupancy as inferred from nucleotide sequence.

DESIGN

Achieving rapid, accurate, and adaptable analysis of CLIP data

The RBPs that shape cellular transcriptomes rely on RNA primary, secondary, and tertiary structure as well as subcellular localization and co-complexes with other RBPs to bind their target sites, but most CLIP studies exclusively evaluate primary sequence enrichment, and few incorporate information from non-uniquely mapping reads that can contain structural motifs for multicopy sequences such as Y RNA, tRNA, and G-quadruplex-containing targets. Furthermore, existing tools that call RNA-protein interactions from CLIP data seldom provide annotations for candidate binding sites aside from overlapping transcript accessions. With Skipper, we implement automated annotation of transcriptomic regions and demonstrate performance across CLIPs for diverse RBPs.

Existing tools that discover RNA-protein interactions from CLIP data often exhibit long runtimes. To enable a major speedup, Skipper first tiles windows over annotated transcripts to create fixed bins for efficiently aggregating read start signal across samples before processing any CLIP read data. Skipper optionally filters out genes that are not expressed in the cell type of interest to improve the accuracy of annotating overlapping features such as coding sequences, introns, or splice sites (Figure 1A). For any customizable set of transcript annotations, Skipper iterates over ranked features and transcript types to create variable-length windows that do not traverse exon-intron junctions or gene boundaries. Skipper partitions the transcriptome into <100-nt windows, corresponding to the length of library fragments generated by eCLIP (Figure 1A).

Skipper then begins sequence read processing (Figure 1B). Reads are trimmed and aligned, and multi-mapping reads are retained. Skipper tallies the counts per window for each sample. Furthermore, reads are aggregated by repetitive element to consolidate multi-mapping sequences and permit quantification of binding to repetitive sequences. To test for enrichment in immunoprecipitation (IP) samples over input samples, a beta-binomial model is fit to the data. Overdispersions in read counts and GC bias are learned for transcriptomic windows and repetitive elements separately, and p values are calculated from the fit beta-binomial distributions.

Extensive quality control summaries are generated for the resulting enriched elements. Output includes tables and visualizations of the number of bound genes (Figure 1C), number of enriched transcriptomic windows (Figure 1D), and the concordance between pairs of replicates (Figure 1E). After processing all replicates separately, reproducible enriched elements are ascertained for both transcriptomic windows and repetitive elements by selecting windows that pass a 20% false discovery in two or more replicates per experiment. Reproducible enriched elements undergo additional, optional, and customizable visualization and analysis as part of a Snakemake workflow[17] (Figure S1; Document S1).

RESULTS

Evaluating candidate binding site detection for diverse RBPs

We find that Skipper's tiled window approach enables efficient analysis of CLIP datasets. Skipper's total runtime is approximately 8-fold reduced compared with our previous peak calling pipeline based on CLIPper[18] (Figure S2A; Table S1). Overdispersion and GC content vary across replicates of eCLIP experiments and complicate significance testing (Figures S2B and S2C). In some cases, bias correction radically alters the calling of enriched windows (Figure S2D). Because many RBPs favor AU-rich or GC-rich sequences,[18,19] GC content is typically confounded with true signal, but correction for ~100-nt windows does not preclude short GC-rich motif enrichment.

To gain insight into Skipper output for diverse RBP binding profiles, we ran Skipper on all eCLIP fastqs available on the ENCODE project website.[18] Across 219 eCLIP datasets, Skipper called an average of 21,310 reproducible enriched windows that serve as candidate binding sites. We compared Skipper's enriched window output with results from running Piranha[6] on the same windows and overlapping reported CLIPper peaks. For 72% of RBPs, Skipper detected more enriched windows than both Piranha (average of 5,039 windows) and CLIPper (average of 6,904 windows). The disparity between Skipper and the other methods was greater for mRNA-binding RBPs than non-coding RNA-binding RBPs (Figure 2A, Table S2).

Whether measured by number of enriched windows (Figure 2B) or agreement between observed enrichment values (Figure 2C), enriched windows detected by Skipper resembled CLIPper output. Skipper rarely increased the number of targets for RBPs that bind small noncoding RNAs such as transfer RNAs (tRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), and Y RNAs: they are usually few in number but with robust IP enrichment.[5] Effect sizes between the two methods were typically correlated around R = 0.6. RBPs that preferred mitochondrial transcripts, snoRNAs, and 5′ UTRs were slightly more consistent (Pearson correlations of 0.77, 0.65, and 0.63, respectively) whereas splice sites and tRNAs or snRNAs (Pearson correlations of 0.50 and 0.51) were more divergent (Figure 2C).

We noticed that some identified candidate binding sites were shared across many RBPs. To mitigate the potential for false positives due to biases in sequence alignment or library preparation, we derived a blacklist from the ENCODE project eCLIP data and filtered out the most common candidate binding sites. RBPs that principally bind small noncoding RNAs (including SBDS, AARS, and SMNDC1) were most affected by filtering: most Skipper-enriched windows were removed and only a small number of snoRNA, tRNA, and snRNA windows remained (Figure 2D). By contrast, for intron and splice site binding proteins, only a small fraction of candidate binding sites were removed.

CellPress
OPEN ACCESS

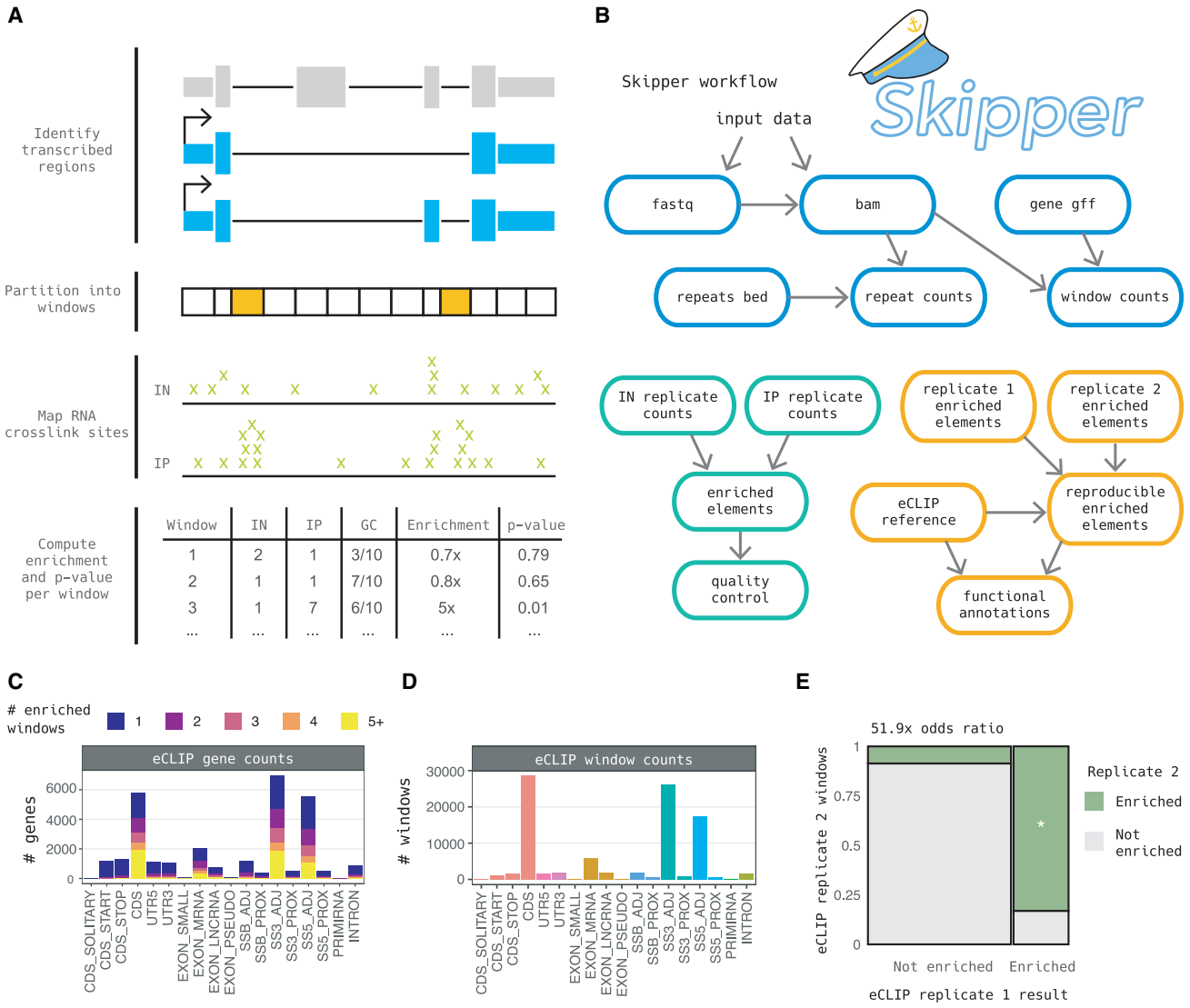


Figure 1. Calling RNA-protein interactions with Skipper

(A) Illustration of Skipper binding calls. Unexpressed genes (gray) are removed. The transcriptome is with <100-nt windows. Crosslink sites (lime) are tallied for input and immunoprecipitated samples. Bins are stratified by GC decile. Significant windows (orange) are detected by beta-binomial testing.

(B–E) (B) Outline of the full Skipper workflow, from fastqs (blue) to enriched loci (green) to annotated reproducible loci (orange). Example output from running Skipper on AQR eCLIP in HepG2 including (C) number of bound genes, (D) number of enriched windows, and (E) concordance between replicates.

Blacklisting eliminates putatively uninformative candidate binding sites that do not depend on the identity of the RNA-binding protein but may also interfere with quantification of binding to small noncoding RNAs.

Skipper outperforms competing methods for calling RNA-protein interactions

Five RBPs with distinct molecular and cellular features were selected for more rigorous evaluation of candidate binding sites: FASTKD2, which binds coding sequences; PUM2, which binds 3′ UTRs; PRPF8, which binds splice sites; TARDBP (or TDP-43), which binds introns; and TROVE2, which binds Y RNAs. Annotations orthogonal to eCLIP data were further used to ascertain true binding sites, specifically mitochondrial transcript identity for FASTKD2, biochemically inferred binding affinity for PUM2,[20] annotated 5′ splice sites for PRPF8, presence of the GURUG motif for TARDBP,[18,19] and Y RNA transcript identity for TROVE2 (Figure 2E).[21]

The number of ascertained true-positive sites called by each method varied drastically across RBPs (Figure 2F; Table S3). PUM2 (6,801 Skipper candidate binding sites in 3,086 genes) was the most consistent across the three methods: nearly half of detected true binding sites were called by Skipper, CLIPper, and Piranha. For PRPF8 (100,581 Skipper candidate binding sites in 8,259 genes), however, Skipper called seven times as many sites as CLIPper and ten times as many sites as Piranha.

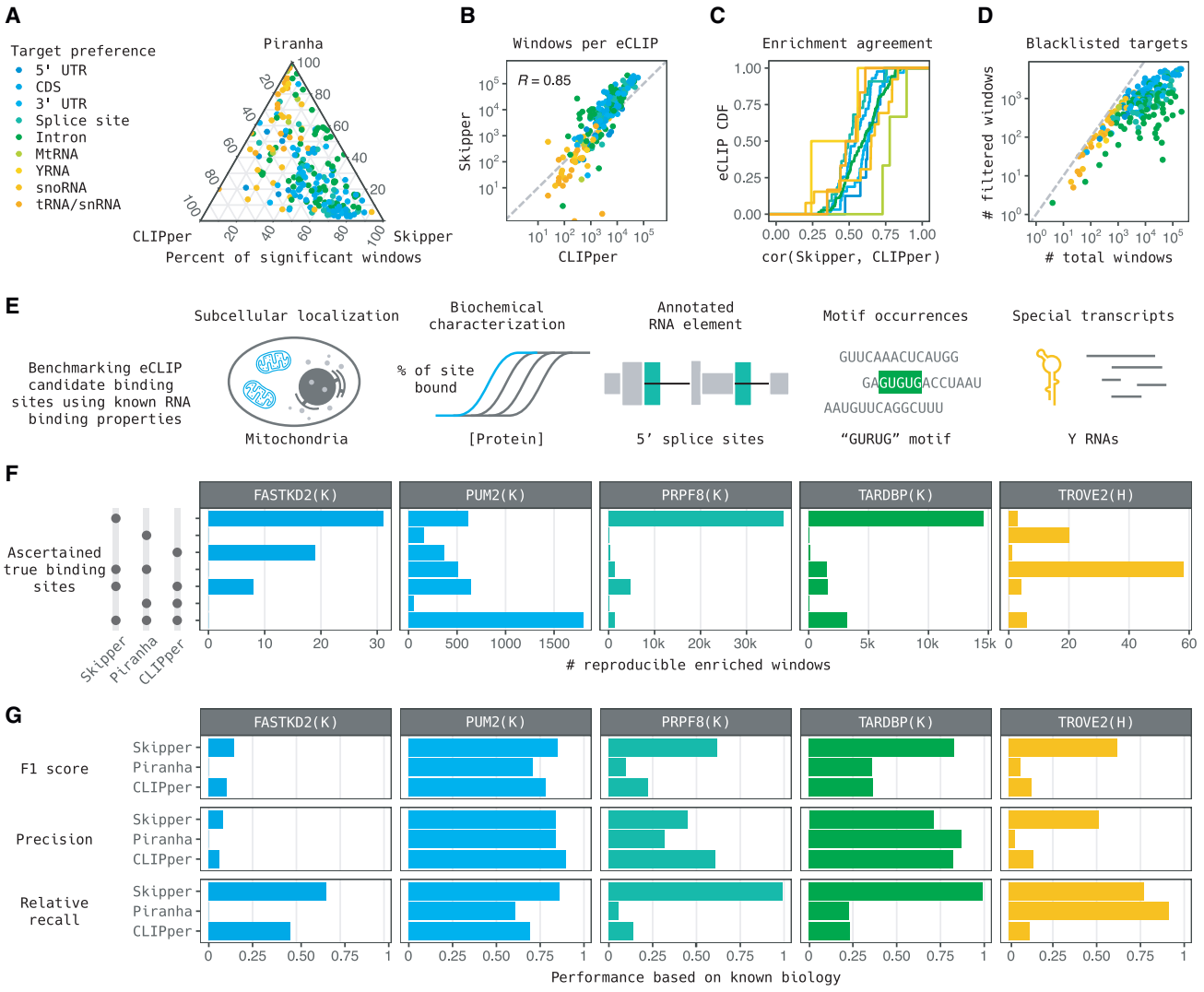


Figure 2. Benchmarking Skipper candidate binding sites

(A) Ternary plot of the proportion of enriched windows called by CLIPper, Piranha, or Skipper from ENCODE eCLIP data.

(B) ENCODE eCLIP experiments visualized by number of bound windows as called by CLIPper (x axis) and Skipper (y axis).

(C) Empirical CDF of the agreement in enrichment values between CLIPper and Skipper as measured by Pearson correlation, stratified by target preference of the RNA-binding protein.

(D) Number of blacklisted nonspecifically bound windows per eCLIP.

(E) Ascertainment of true binding site windows for example eCLIPs.

(F) Counts of true binding site windows per example eCLIP.

(G) Example eCLIP precision, recall, and F1 score per method.

TARDBP (29,330 Skipper candidate binding sites in 5,550 genes) exhibited a similar, albeit attenuated, trend: three times more sites with Skipper than with CLIPper or Piranha. In the case of FASTKD2 (486 Skipper candidate binding sites in 251 genes), mitochondrial transcripts were blacklisted and contributed zero candidate binding sites under Piranha's algorithm. TROVE2 (138 Skipper candidate binding sites in 63 genes) was the only eCLIP dataset for which Skipper did not call the most candidate binding sites.

Although Skipper reported the greatest number of ascertained true binding sites, the purity of true positives among called binding sites (the precision) is essential for judging performance. Contrary to our expectation that methods with higher sensitivity would attain lower specificity, the three methods exhibited similar precision across large differences in the number of sites called (Figure 2G). Skipper attained the highest precision for FASTKD2 and TROVE2, Piranha for TARDBP, and CLIPper for PUM2 and PRPF8. Three cases exhibited very low precision: Piranha and CLIPper on TROVE2 and Piranha on FASTKD2 due to blacklisting of mitochondrial transcripts. Increasing the stringency for reproducible windows slightly improved the precision of Skipper calls but gravely reduced recall (Figure S2E).

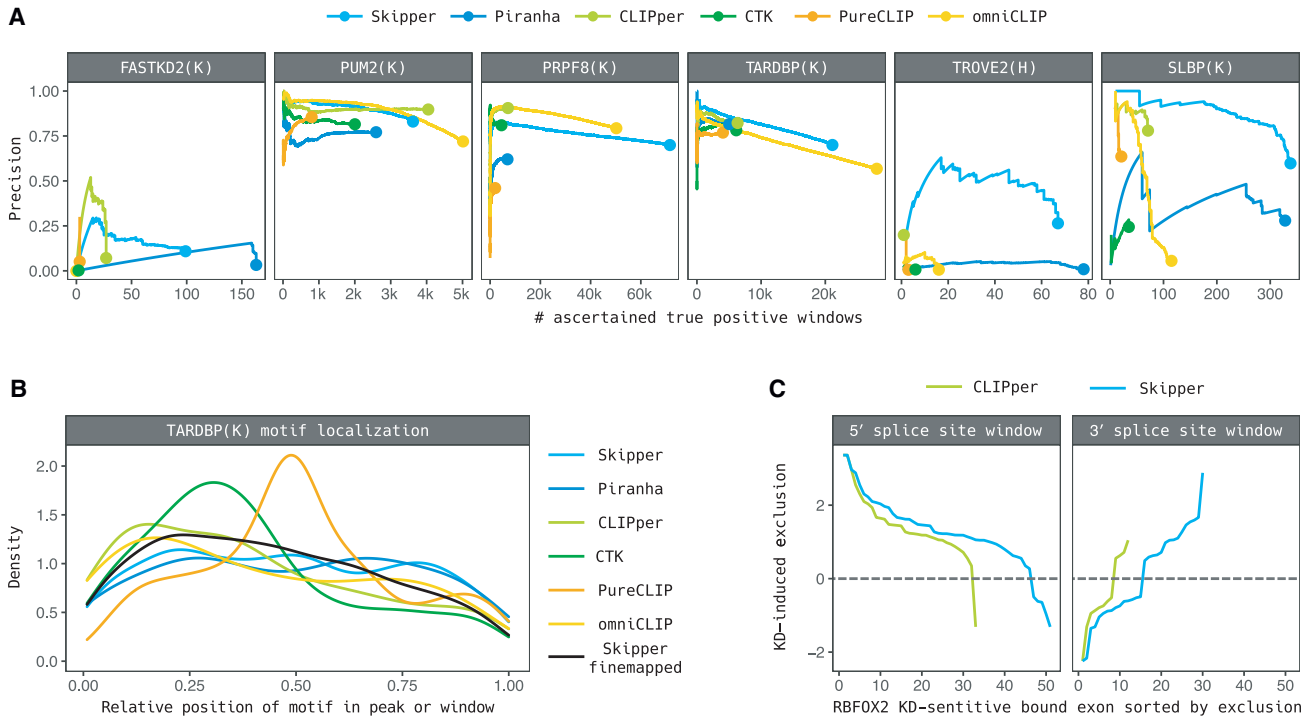


Figure 3. Comparing CLIP tool performance

(A) Precision-recall curves for Skipper, Piranha, CLIPper, CTK, PureCLIP, and omniCLIP for example CLIPs.

(B) Relative density of the distance from predicted crosslink sites, peak, or window centers to TDP-43 (TARDBP) motif occurrences for the above methods.

(C) Knockdown-sensitive RBFOX2 exons sorted (x axis) by average reported SepScore (y axis) for proximity to both 5′ (left) and 3′ (right) splice sites.

Having established competitive levels of precision for Skipper, we assessed Skipper candidate binding sites via other metrics. We calculated the percentage of binding sites detected out of the union of ascertained true binding sites across all methods (the relative recall) and the harmonic mean of the precision and relative recall (the F1 score) for each eCLIP (Figure 2G). Skipper attained the highest F1 score across all examples. CLIPper output for FASTKD2 and PUM2 were the only cases where Skipper's improvement did not exceed other methods' F1 score by more than 10%. We next investigated how CLIPper, Skipper, and Piranha candidate binding sites varied across transcript abundance levels by binning transcripts and calculating the average probability of enrichment across all eCLIPs per bin. For CLIPper and Skipper, windows tiling the top decile of expressed transcripts were more than twice as likely to be candidate binding sites on average (Figure S3A). Piranha exhibited even more extreme bias: the majority of all candidate binding sites are in transcripts in the top decile of expression.

Because Skipper attained the greatest improvement in performance for TROVE2 binding Y RNAs, we also evaluated eCLIP of SLBP (439 Skipper candidate binding sites in 114 genes), which binds another unique class of transcripts: histone mRNAs harboring stem loops. Precision was comparable for all three methods, but Skipper again demonstrated the greatest F1 score (Figures S3B and S3C).

For more context, we investigated CLIP analysis tools that employ approaches distinct from Skipper's, CLIPper's, or Piranha's. We processed the six example eCLIP datasets above with CLIP Tool Kit (CTK),[22] which calls local maxima in CLIP read signal tracks, and PureCLIP[9] and omniCLIP,[10] which use a nucleotide-resolution hidden Markov model to segment the transcriptome into no read signal, background read signal, and bound sites. Both omniCLIP and PureCLIP use a generalized linear model to adjust emission probabilities based on background read signal; however, although PureCLIP uses a zero-truncated binomial model of read starts to discriminate between authentic and inauthentic binding signals, omniCLIP uses a Dirichlet-multinomial distribution of varied diagnostic events (read starts, insertions, and all possible mismatches).

Skipper attained comparable or superior performance across all example eCLIPs and tools (Figure 3A; Table S4). Blacklisting was not performed to put all tools on equal footing. For all examples, CTK performed strictly worse than CLIPper, and PureCLIP identified very few ascertained true-positive windows. omniCLIP appeared to perform marginally better than Skipper for PUM2 and PRPF8, and marginally worse for TARDBP. However, omniCLIP reported zero mitochondrial windows for FASTKD2 and predominantly reported false positives for TROVE2 and SLBP. FASTKD2, TROVE2, and SLBP possess a small number of ascertained true-positive windows in the transcriptome, which may pose a particular challenge for approaches that assess the likelihood of the whole dataset.

Curious about the differences across tools for detecting true-positive windows, we further examined tool output. We

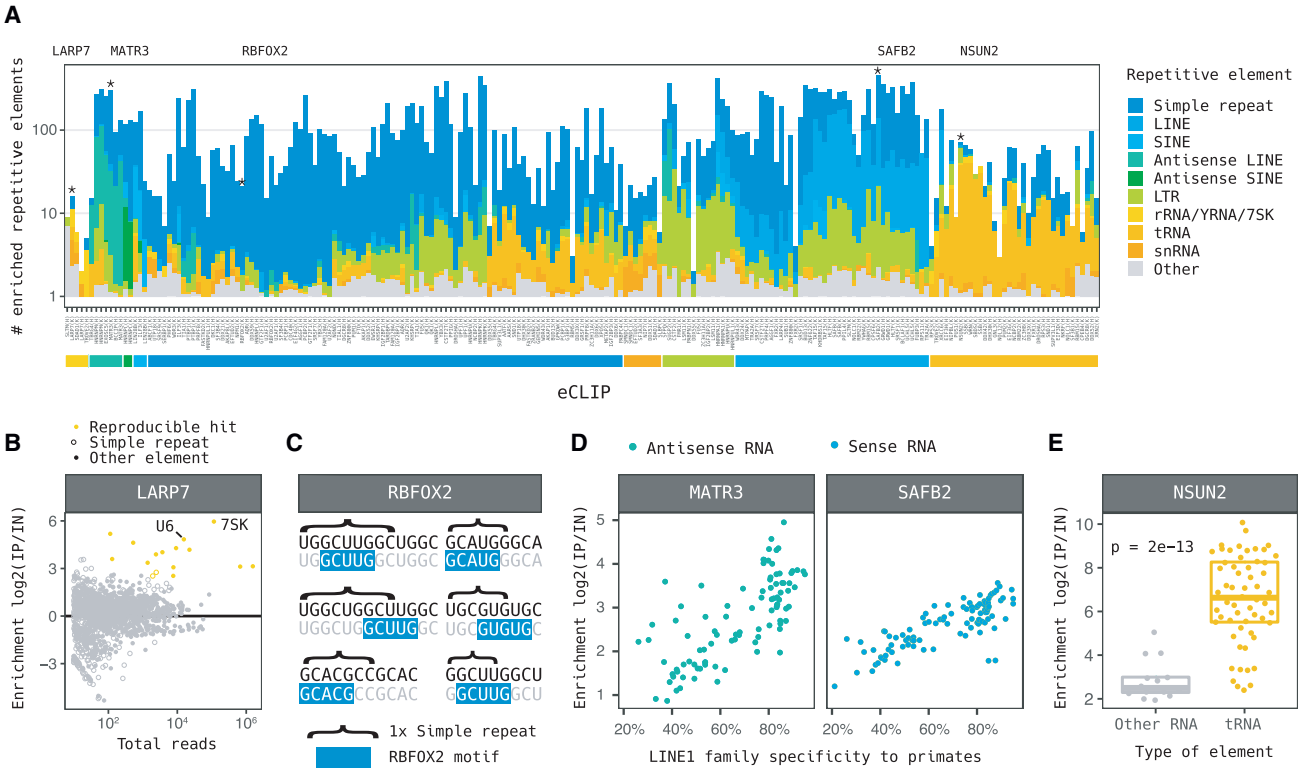


Figure 4. Skipper quantification of repetitive element binding

(A) Counts of enriched repetitive elements per CLIP experiment, colored by type of repetitive element. Asterisks denote LARP7, RBFOX2, MATR3, SAFB2, and NSUN2, which are visualized in greater detail.

(B) LARP7 enrichments per repetitive element.

(C) The top six RBFOX2 binding motifs within enriched simple repeats.

(D) Skipper binding enrichment plotted against evolutionary age of LINE1 sequences for MATR3 (left) and SAFB2 (right), antisense in green and sense in blue.

(E) Repetitive element targets of NSUN2. tRNA targets exhibit greater enrichment than other RNAs (p = 2e–13)

visualized the density of TARDBP GUGUG motif occurrences within peaks or windows containing one GUGUG occurrence. Because Skipper windows are tiled in advance irrespective of read signal, we implemented a fine-mapping approach to center 75-nt windows on local maxima enrichment within Skipper windows, which we confirmed increased GURUG motif density 1.4-fold for TARDBP in K562 (t test p = 5e−138, N = 18,508 windows) and GCAUG motif density 1.3-fold for RBFOX2 in HepG2 (t test p = 1e−123, N = 15,794 windows).

We found that the maximum motif density of each tool appeared to be anticorrelated with true-positive window F1 scores (Figure 3B). PureCLIP exhibited the most focal signal: identified crosslink sites exhibited maximum motif density more than double the average. CTK exhibited the next greatest localization of motif density, but the maximum density was slightly upstream of the center of called peaks. omniCLIP, CLIPper, and Skipper after fine mapping all exhibited a mild increase in motif density upstream of inferred crosslink sites. Piranha and Skipper before fine mapping exhibited flat distributions. From this, we posit that less sensitive tools more precisely register foci of enrichment but struggle to detect broad stretches of more weakly enriched binding sites that may contain secondary binding motifs. Examination of binding near alternative exons sensitive to knockdown also revealed that Skipper detected more RBFOX2 candidate binding sites flanking knockdown-sensitive alternative exons than CLIPper (Figure 3C).

A compendium of repetitive element binding

Our past work mapped reads to repetitive elements and examined information content across eCLIPs but did not report bound elements and limited its investigation of repetitive element enrichment to 65 eCLIP datasets.[12] With Skipper, we identify bound repetitive elements in eCLIP data using the same overdispersion and GC-content modeling framework. Nearly all (216 out of 219) eCLIP datasets exhibited enrichment of repetitive elements (Figure 4A; Table S5). Repetitive element-binding proteins achieved high specificity for known RNA templates, such as LARP7 to its targets 7SK and U6 (Figure 4B).

Across all RBPs, the most frequently bound class of repetitive elements were simple repeats of 1–12 nt. Although we previously reported RBP motifs from reads transcriptome-wide, it was not established whether long motif repeats show comparable specificity.[18] We clustered mono-, di-, and tri-nucleotide repeat-binding profiles and found both broad- and fine-scale patterns of selectivity (Figure S4). GU-rich sequences were the most frequently bound repeats, but some RBPs (e.g., LIN28B and

LSM11) bound (UGG)n and not (GU)n repeats, whereas others (e.g., SRSF7 and TIAL1) bound (GU)n and not (UGG)n repeats. Even when RBPs bound both (UGG)n and (GU)n repeats, they differed in whether they bound (GUU)n repeats (e.g., HNRNPM and SUBP2), (AUG)n repeats (e.g., EFTUD2 and FAM120A), or neither (e.g., NCBP2 and FUS).

Some repeats exhibited nuanced specificity for groups of similar k-mers. Inspection of RBFOX2 simple repeats binding in K562 cells revealed that all six of the top simple repeats contained one of eight established motifs directing canonical binding[23] (Figure 4C), and, out of all 16 simple repeats called, 11 contained one. QKI exhibited very high enrichment for (UAC)n repeats. QKI's canonical binding motif UACUAACN$_{1\text{-}20}$UAAY[24] may commonly arise within annotated (UAC)n repeats (Figure S4). Finally, C homopolymers were bound by a small number of factors known to favor C-rich sequences including HNRNPK and PCBP2.

Other classes of bound repetitive elements generally agreed with known biology. MATR3, which extensively binds antisense LINE1 transcripts,[25] was among the top antisense LINE1 RBPs, and SAFB2, which represses LINE1 transcripts,[26] among sense LINE1 RBPs. Skipper tests binding to annotated sets of repetitive elements and robustly calculates aggregate enrichment regardless of read depth. We found that the reported enrichments strongly correlated with evolutionary age of the LINE1 element for both RBPs, and all bound LINE1 elements were of the expected strand (Figure 4D).[27,28] Antisense SINE elements were bound most by HNRNPC[29] and sense SINE elements by ILF3.[30] We identified three main biological processes governing tRNA interactions: translation (EIF3G, RPS3, and SBDS), RNAi (SND1, DGCR8, and DROSHA), and tRNA modification (NSUN2, PUS1, and SSB). In some cases, quantitative enrichment level strongly implicated the most salient targets even when multiple types of repetitive elements were identified (Figure 4E). Finally, snRNA-binding proteins included splicing regulators such as SMNDC1 and EFTUD2.

A modest number of RBPs principally interacted with long terminal repeats (LTRs), which was not apparent using our previous enrichment filters.[12] Motif analysis with HOMER yielded mostly G- and GU-rich binding motifs. Notably, several of the RBPs in this class contain RGG motifs known to mediate RNA-protein interactions[31]: FUS, HNRNPA1, HNRNPU1, and HNRNPUL1. HNRNPK, which clustered with simple repeat-binding proteins, also contains an RGG motif and bound numerous LTR sequences. RGG domains are thought to mediate binding to G quadruplexes,[32–34] and recent work posits that stabilization of viral LTRs by RGG domain-containing proteins could play an important role in suppressing viral protein expression.[35,36] Thus, in addition to their posited role in viral defense, RGG domain-containing proteins may also guard against reactivation of endogenous retroviruses.

Archetypes of alternative exon binding

Past studies have established enrichment of RBP binding sites at splice sites flanking alternative cassette exons and a depletion of RBP binding sites residing in alternative cassette exons.[18,37] In aggregate, these binding profiles appear to promote intermediate splicing; however, the identity of the bound RBP is critical in determining whether splice site usage is enhanced or silenced. Indeed, mutation of RBP binding sites induce splicing changes in accordance with the corresponding RBP's distinct gene regulatory role.[38–41]

To examine alternative splicing regulation at the level of individual RBPs, we overlapped 5′ and 3′ splice sites flanking alternative cassette exons with the expanded set of enriched windows called by Skipper. We stratified the alternative exons by whether an RBP of interest bound the alternative splice site window, the constitutive splice site window, or both and searched for stereotyped RBP binding patterns (Figure 5A). Four patterns stood out, exemplified by SF3B4, RBM22, BUD13, and HNRNPC (Figure 5B). RBPs such as SF3B4 principally bound 3′ splice site windows and favored constitutive 3′ splice site windows (8,916 enriched splice site windows). RBPs such as RBM22 principally bound 5′ splice site windows and favored both constitutive 5′ splice site windows and alternative 3′ splice site windows (2,403 windows). RBPs such as BUD13 commonly bound both 5′ and 3′ splice site windows with a constitutive bias for both splice site windows (19,100 windows). Finally, RBPs such as HNRNPC bound both 5′ and 3′ splice site windows with mild to moderate alternative window bias (354 windows).

We then inspected the level of inclusion of skipped exons stratified by whether each RBP bound 5′ and/or 3′ splice site windows (Figure 5C). Alternative 3′ splice site binding by SF3B4 appeared to increase exon inclusion. Binding by RBM22 appeared to increase exon inclusion in alternative 5′ splice site windows and decrease exon inclusion in alternative 3′ splice site windows. In contrast, BUD13 binding appeared to increase cassette exon inclusion in the vicinity of either the 5′ or 3′ splice site. Conversely, HNRNPC binding appeared to decrease cassette exon inclusion when binding in the window containing either the 5′ or 3′ splice site.

Across all eCLIP experiments, most RBPs exhibited a binding profile that aligned with one of the four archetypes we described (Figure 5D, Table S6). 3′ splice site regulators including U2AF2 and SF3A3 were represented in the 3′ splice site enhancing archetype. Essential splicing factors such as AQR and PRPF8 comprised the 5′ and 3′ splice site enhancing archetype. Splice modulators such as HNRNPM and SUGP2 belonged to the 5′ and -3′ splice site silencing archetype. RBPs that affect mature mRNA stability, such as DGCR8, PUM1, CPSF6, and UPF1, may be acting on overlapping alternative exons rather than intronic splicing elements.

Members of the 5′ splice site enhancing/3′ splice site silencing archetype are less often associated with regulation of alternative exon abundance (Figure 5D). Several RBPs in this class bind 5′ UTRs (e.g., FTO, DDX3X, and NCBP2) or G- and GU-rich repeats (e.g., FUS, NKRF, and GTF2F1), and we found instances of RBPs that favor 3′ constitutive splice sites without associated changes in exon inclusion (e.g., LSM11 and CSTF2). We conclude that alternative splice site binding preferences may sometimes relate to proximity to UTRs rather than alternative RNA splicing or decay.

We repeated the binding site analysis using alternative 3′ splice sites and found that most significant differences in splice site occupancy reflect exon-binding proteins that bind alternative exons of mature mRNA in the cytosol and not intronic windows in the

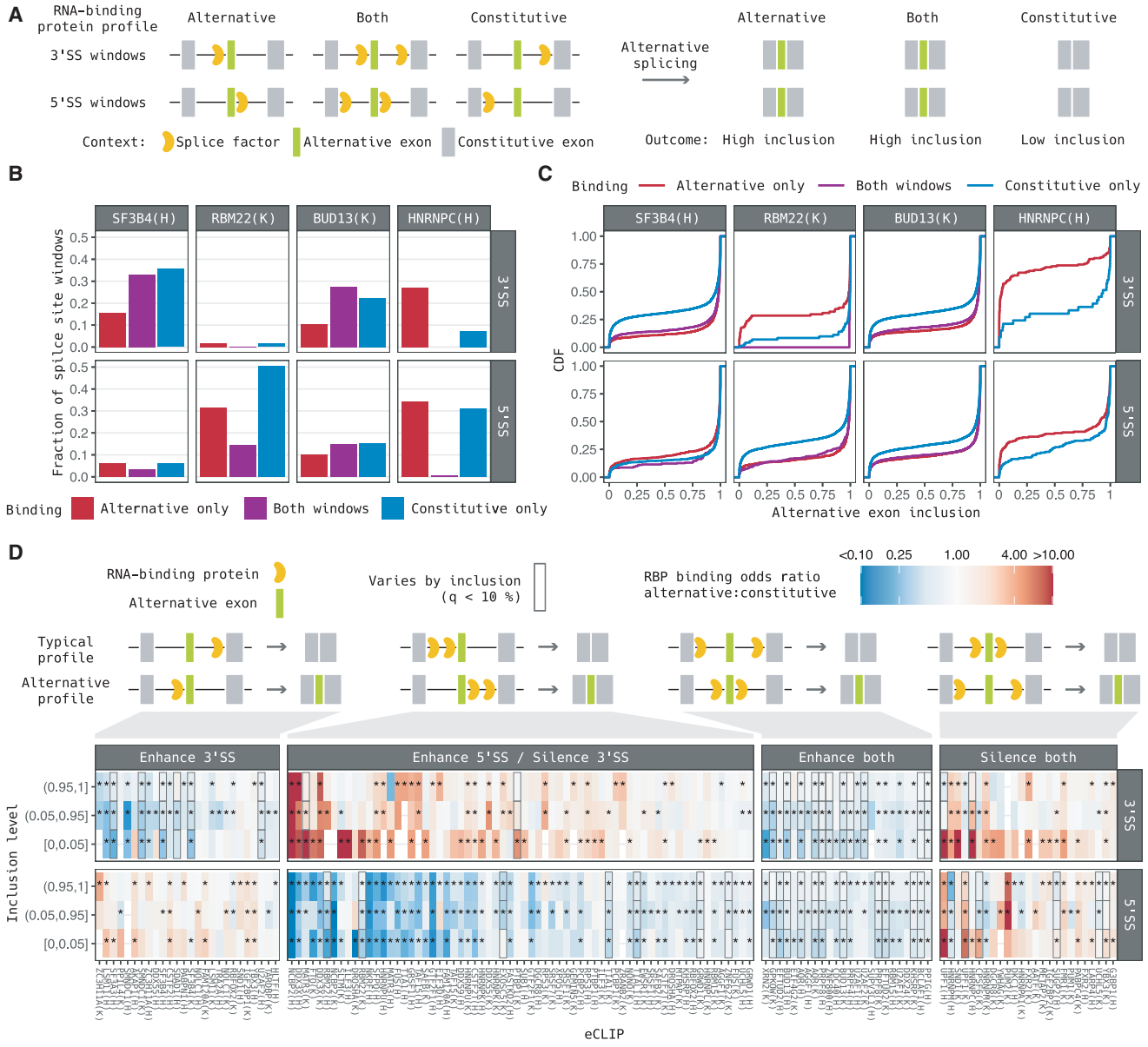


Figure 5. Archetypes of RNA-protein interactions near alternative cassette exons

(A) Schematic of splicing regulator binding of 3′ and 5′ alternative splice site windows.

(B) Frequency of alternative cassette exons bound to alternative (red), constitutive (blue), or both (purple) 5′ and 3′ Skipper splice site windows.

(C) Empirical CDFs of alternative cassette exon inclusion stratified by enriched window overlap with 5′ splice sites, 3′ splice sites, or both.

(D) Typical and alternative binding profiles of the four archetypes associated with skipping (top) and inclusion (bottom), respectively. Heat shows alternative (red) and constitutive (blue) exon-biased RBPs. Significant bias is denoted by asterisks and significant covariance by boxes.

nucleus (Figure S5; Table S7). Even so, splice-repressive RBPs HNRNPL and HNRNPK silenced splicing in both alternative and constitutive windows, and splicing RBPs AQR and PRPF8 promoted splicing in both alternative and constitutive windows. RBPs associated with alternative splicing, including RBFOX2, EFTUD2, and SMNDC1, bound the window flanking the alternative, longer exon regardless of inclusion level.

We crosschecked the 31 RBPs for which binding correlated with isoform abundance against RBPs known to regulate RNA splicing or decay. Members of our list of candidates were three times more likely to be known RNA splicing factors than other RBPs that bound splice site windows ($p = 0.01$, $N = 100$ total RBPs, Fisher's exact test).[42–44] Four other RBPs (YBX3, EXOSC5, UPF1, and MATR3) were annotated as playing a role in post-transcriptional gene regulation.

After removing the RBPs known to contribute to post-transcriptional regulation of gene expression, seven candidates for unannotated regulation of RNA splicing and decay remained:

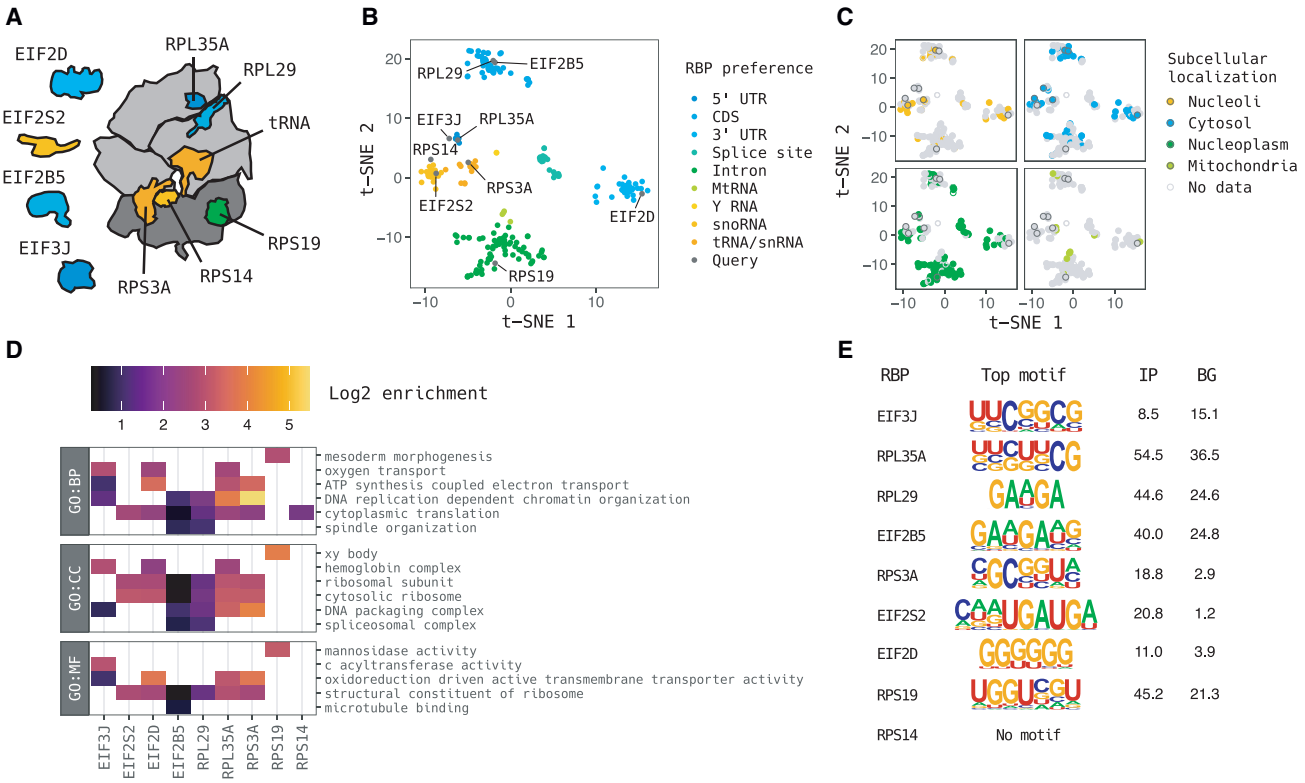


Figure 6. Determinants of translation factor occupancy

(A) Illustration of profiled translation factors.

(B) Skipper t-SNE query of translation factors against ENCODE eCLIP data colored by target preference.

(C) Skipper t-SNE points recolored by subcellular localization: nucleoli (orange), cytosol (blue), nucleoplasm (dark green), and mitochondria (lime).

(D) Significant Skipper Gene Ontology enrichments for translation factors filled by $\log_2$ enrichment. The top cellular component, biological process, and molecular function term for each eCLIP is shown.

(E) HOMER motifs for each translation factor.

UCHL5, binding of which was recently linked to splicing changes via a multiplex splicing minigene screen[45]; GRWD1, binding of which is often disrupted by splicing quantitative trait loci (sQTLs)[38]; and ZNF622, ZNF800, BCLAF1, DDX24, and SDAD1, for which binding has not been associated with RNA splicing or decay. ZNF622, ZNF800, BCLAF1, YBX3, and GRWD1 are putative transcriptional regulators that could influence nascent RNA processing.[46] DDX24 has been linked to RNA trafficking,[47] and SDAD1 is almost entirely unannotated.[48] Notably, review of the repetitive element targets of UCHL5, GRWD1, ZNF622, ZNF800, BCLAF1, and DDX24 revealed primarily sense LINE1 transcripts (odds ratio [OR] = 14 for the candidates over all other RBPs, p = 9e−5, Fisher's exact test, Figure 4A), suggesting that these factors may alter excision of introns containing LINEs.

eCLIP of translation factors captures diverse molecular interactions

To demonstrate application of the Skipper pipeline, we generated eCLIP data on a batch of translation factors with validated IP-grade antibodies: EIF2D, EIF2S2, EIF2B5, EIF3J, RPL35A, RPL29, RPS3A, RPS14, and RPS19 (Figure 6A). The binding preferences reported by Skipper varied widely: RPL29 and EIF2B5 favored coding sequences, EIF3J and RPL35A favored 5′ UTRs, RPS14 and EIF2S2 favored snoRNAs, RPS3A favored tRNAs, EIF2D favored 3′ UTRs, and RPS19 favored introns (Figure 6B).

To interrogate the binding preferences of the selected translation factors further, we annotated the RBPs with their observed protein subcellular localization[49] (Figure 6C). RPS19 localized principally to the nucleoplasm, consistent with its strong intronic binding preference. EIF2S2 and RPS3A were detected in nucleoli, consistent with their preference for binding snoRNAs and tRNAs. The other translation factors localized to the cytosol, consistent with mRNA binding.

To discern which gene pathways were most enriched for translation factor binding, we performed a weighted Gene Ontology enrichment that tallied the number of enriched windows per term aggregating across the transcriptome (Figure 6D; Table S8). Among the translation factors we profiled, the most frequently enriched term was "cytoplasmic translation." Related top terms include "structural constituent of the ribosome," "cytosolic ribosome," and "ribosomal subunit." Two other groups of genes were highly enriched: histone genes ("DNA packaging complex" and "DNA replication dependent chromatin organization") and respiration ("ATP synthesis coupled

electron transport'' and ''oxidoreduction driven active transmembrane transporter activity'').

We next ran HOMER to call sequence motifs underlying RBP binding. EIF2S2's fine-mapped windows returned the snoRNA C box motif UGAUGA,[50] consistent with its snoRNA binding preference. RPL35A and EIF3J, both selective for 5′ UTR, exhibited motifs that reflect KYCKKCG binding. EIF2B5 and RPL29, both selective for coding sequences, both exhibited purine-rich motifs. RPS19 yielded a GU-rich motif similar to other intron-binding proteins: UGGUNGU. RPS3A was associated with GCGGU sequences, and EIF2D, known to regulate ribosome recycling near stop codons,[51] was associated with G-rich sequences in 3′ UTRs. RPS14 yielded no apparent motif.

Despite comprising many of the same protein complexes, different translation factors interacted with distinct sites in the human transcriptome. Binding profiles appeared distinct both from each other (Figure S6A) and from published ENCODE binding profiles (Figure S6B). Thus, translation factor binding profiles can reflect distinct biological processes that occur in varied subcellular compartments and exhibit distinct sequence preferences.

Depletion of genetic variation nominates transcripts with constrained binding

Given the diverse sequence preferences of our panel of translation factors, we wondered whether perturbation of occupied sites would interfere with regulation of translation. To test for constraint acting on sequence-driven translation factor occupancy, we trained a gapped k-mer support vector machine (gkm-SVM) on fixed 75-nt windows centered on signal overlapping enriched windows from Skipper and evaluated whether disruptive genetic variants were depleted in the gnomAD genetic database.[37,38,52–54]

gkm-SVMs for all translation factors but RPS14 exhibited significant separation between bound and control sites (Figure S7A). EIF2B5 and RPS19 target sites attained the greatest performance (area under the precision-recall curve [AUCPR] of 0.691 and 0.625), while the others exhibited low to moderate performance (AUCPR from 0.21 to 0.429). We moved forward with assessing potential constraint for all models but RPS14's.

We queried gnomAD for genetic variants in fine-mapped windows for each RBP and binned variants by allele frequency: singleton, very rare (<0.1%), rare (0.1%–1%) or common (>1%) (Figure 7A). Reference and variant 75-nt sequences corresponding to Skipper fine-mapped windows were scored by the corresponding gkm-SVM to yield a delta score representing the predicted change in binding from the reference to the variant sequence. Delta scores were fit per transcript using linear regression against variant frequency bin as an ordinal variable. We interpreted greater slope in the linear regression as reflecting greater selective constraint on translation factor occupancy.

Some significant translation factor-transcript pairs were especially intriguing. Constrained RPS19 binding to *LAMP1* and *MAN1B1* transcripts was driven by multiple intronic windows. Singleton variants in fine-mapped windows had far lower delta scores than the transcriptome-wide average, whereas variants above 0.1% frequency had far higher delta scores (Figure 7B, upper boxplots). The same trend was observed for EIF2B5 binding to coding sequences of *GRSF1* and *RPL5*, even without a transcriptome-wide trend disfavoring variants with lower delta scores (Figure 7B, lower boxplots).

Overall, we detected 65 transcripts that exhibited constrained binding by a translation factor under a 10% FDR (Figure 7C; Table S9). Hits included long noncoding RNAs (e.g., *Chaserr*), demonstrating potential for discovery of functional noncoding binding sites even for translation-associated proteins. For RBPs that principally bound mRNAs, we also investigated characteristics that increased the likelihood of calling constrained transcript binding. RPL29, EIF2B5, EIF3J, and RPL35A were more likely to exhibit constrained binding in 5′ UTRs and less likely in introns and coding sequences (Figure S7B). By contrast, transcripts with reproducible enriched windows for RPS19 were much more likely to be constrained for intronic windows than coding windows. Thus, our nominated constrained binding events appear to recognize different gene regulatory roles acting on different transcript regions.

In order to link changes in RPS19 binding to altered risk for Diamond-Blackfan anemia, we exogenously expressed wild-type and mutant RPS19 constructs in HEK293T cells and performed CLIP for the RPS19 missense mutations most common among Diamond-Blackfan anemia cases: dominant-negative mutations R62W and R101H (Figure 7D).[55–57] Large-scale changes in RNA interactomes were observed for both mutant RPS19 constructs: an increase in the proportion of coding sequences (CDSs) from 33% to 45% and a decrease in the proportion of deep introns from 20% to 7% ($p < 2.2e{-}16$, chi-square test; Figure 7E) and greater enrichment of tRNA targets (R62W $p = 6e{-}9$, R101H $p = 2e{-}27$, N = 61 tRNA genes, paired t test; Figure 7F). Other elements annotated by RepeatMasker did not show enrichment in the mutant CLIP samples (Figure S7C). To better understand global changes in RPS19 binding induced by Diamond-Blackfan anemia mutations, we aggregated signal gene-by-gene for deep introns and CDSs separately and correlated signal across all pairs of replicates. The mutant RPS19 CLIP replicates exhibited highly correlated binding profiles ($R > 0.95$) with each other, for both CDSs and deep introns (Figure S7D). However, the divergence of mutant from wild-type RPS19 interactions was much more pronounced for deep intronic binding ($R \sim 0.5$) than for CDS binding ($R \sim 0.85$) (Figure S7D). Overall, the two Diamond-Blackfan anemia mutations induced similar changes in CLIP signal, which suggests a shared mechanism for disruption of erythroid development.

DISCUSSION

Our CLIP-seq processing tool Skipper offers fast, customizable, and comprehensive analysis of CLIP-seq data by assessing read starts in windows for IP versus matched input samples. Skipper matches or exceeds previous methods in precision and calls more than three times as many candidate binding sites from eCLIP data available on the ENCODE project portal. With respect to interpretation of post-transcriptional functional consequences of genetic variation, Skipper increases the number of overlapping windows 2.5-fold for GTEx lead eQTLs and sQTLs (Figure S7E, left two panels), and 4-fold for lead 3′ UTR alternative

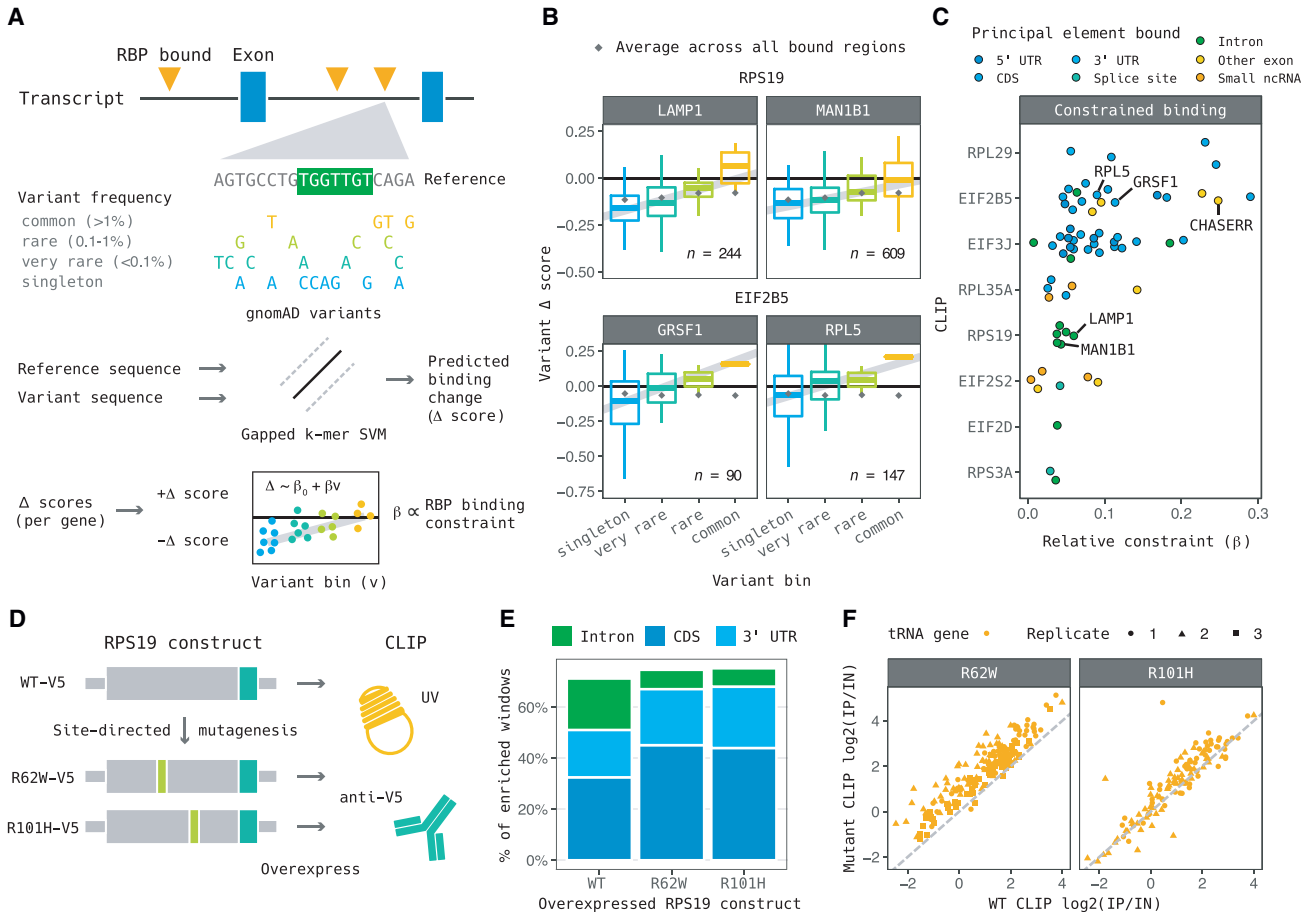


Figure 7. Identification of sequence-constrained translation factor binding

(A) Schematic of constraint detection procedure from gnomAD variants to linear regression of predicted binding change against allele frequency bin.

(B) Transcripts with apparent sequence constraint bound by RPS19 and EIF2B5.

(C) The 65 RBP-transcript constrained pairs detected colored by the principal feature bound.

(D) Schematic of RPS19 CLIP experiments in HEK293T cells.

(E) Percentage of enriched windows deriving from coding sequences, 3′ UTRs, and introns for wild-type and mutant RPS19 CLIPs.

(F) Enrichment for tRNA transcripts in mutant (y axis) versus wild-type (x axis) CLIP experiments.

polyadenylation QTLs (3′ aQTLs) (Figure S7E, right two panels). Unlike existing methods, Skipper also aggregates reads across instances of repetitive RNA elements and reports statistically enriched elements. Skipper's automated visualizations expedite quality control of CLIP-seq data, and the corresponding tabular output provides a launching pad for exploration of RNA-protein interactions in high throughput.

Evaluation of CLIP-seq data quality has overwhelmingly focused on detecting and quantifying motif binding,[6–9,58] but this approach overlooks many known determinants and consequences of RNA-protein interactions: subcellular localization that physically separates transcripts from RBPs, differences in binding affinity that defy position weight matrix predictions, essential regulatory proteins such as ribosome subunits that can translocate along RNA in a motif-independent manner, and specialized modifications or conformations that proteins recognize independent of sequence. We found that use of a size-matched input control improves performance dramatically in the case of the last group: in particular, Y RNAs by TROVE2 (Figure 2G), tRNAs by NSUN2 (Figure 4E), and histone mRNAs by SLBP (Figures S3B and S3C).[5] The relatively small number of authentic targets for these RBPs increases vulnerability of analyses to false positives.

Similarly, the value of Skipper repetitive element output should not be overlooked. Binding to simple repeats was widespread (89.5% of RBPs exhibited 10-fold enrichment in binding to one or more simple repeats) but also distinct for each RBP (59.4% of RBPs possessed at least one simple repeat that no other RBP bound). Thus, our compendium should aid future efforts to use decoy RNAs[59] to sequester RBPs at variable levels of specificity for diverse regulatory RBPs. Conversely, evaluation of potential decoy RBPs[60] to block binding sites should be greatly facilitated using Skipper's assignment of simple repeat preferences.

Investigation of RBP binding near alternative exons offers a complementary approach to post-transcriptional regulation.

Characterizing RBP function using gene knockouts is often made impractical by the broad essentiality of RBP-regulated processes. While integration of eCLIP data with knockdown RNA-seq can define a causal link between binding and regulatory outcomes, siRNA knockdown is prone to off-target effects,[61] and imperfect calls for both RBP binding and isoform quantification (e.g., from GC bias and insufficient coverage) reduce the number of intersecting sites.[18] By correlating binding signal with alternative exon inclusion, we can infer many of the same relationships between RBP occupancy and RNA processing without collecting an independent knockdown RNA-seq dataset (Figure 5D). Our association of YBX3, ZNF622, ZNF800, BCLAF1, and GRWD1 differential binding to changes in isoform abundance suggests that the role of putative chromatin modulators in regulating RNA splicing and decay may be an underrecognized source of co- or post-transcriptional gene regulation.

Finally, our approach for detecting constraint acting on binding sites at the level of individual transcripts offers a roadmap for probing different layers of regulation by RBPs. *Chaserr*, an essential long non-coding RNA (lncRNA), serves as an example. Regulatory mechanisms for lncRNAs occur at myriad levels in gene regulatory networks from enhancer competition to post-translational modifications.[62] *Chaserr* is thought to regulate its neighboring gene CHD2 in *cis*, but the precise ways in which the *CHASERR* gene body fulfills its gene regulatory role are unknown.[63] Selective constraint on sequences in loci occupied by translation factors suggests novel molecular mechanisms for regulation by *Chaserr* at the level of translation. The diverse types of transcripts, subcellular localizations, motifs, and RNA regions we identified under constraint per translation factor occupancy (Figures 6 and S6) reinforce the broad applicability of our approach.

Testing for transcript-level constraint can also nominate mechanisms responsible for human disease. Mutations in RPS19 are the most common single-gene cause of Diamond-Blackfan anemia. The molecular pathophysiology of Diamond-Blackfan anemia entails defective rRNA maturation, reduced translation of GATA1,[64] and impaired erythroid development, but why the mutations disproportionately occur in RPS19 has not been established.[55] Furthermore, patients exhibit considerable heterogeneity in clinical presentation.[65] Enrichment of disease variants occurring in RNA-binding protein binding sites is well documented, yet such work usually aggregates signal at the level of whole transcriptomes.[34,39,53] By contrast, the transcripts with sequence-dependent RPS19 intron binding we identified under selective constraint could point to pathways that underlie erythroid susceptibility to mutations found in Diamond-Blackfan anemia patients as well as differing clinical presentations.

Our results point to multiple grounds for future investigation. The transcript with the most constrained RPS19 binding, LAMP1 (Figure 7C), could conceivably play a role in Diamond-Black anemia pathophysiology. LAMP1 is one of the main constituents of lysosomal membranes. Erythroblasts rely on autophagy via lysosomes to eliminate organelles that impede erythrocyte maturation and function.[66] One study found that knockout of factors essential for autophagy causes anemia in mice,[67] and an unbiased chemical screen revealed that induction of autophagy with the small molecule SMER28 enhanced erythropoiesis in induced pluripotent stem cells derived from Diamond-Blackfan anemia patients.[68] Previous work has shown that RPS19 R62W persistently localizes to the nucleus[56] but alters cellular morphology even at moderate expression levels.[57] Deeper insight into the link between perturbed interactions with RNA and altered cellular morphology could illuminate the mechanism underlying the apparent dominant-negative mode of inheritance.

As methodological and computational approaches to cataloging RNA-protein interactions continue to improve, understanding the functional significance of RBP binding will grow only more important. Determining whether an individual RBP binding site is under selective constraint remains challenging because a genetic variant's predicted change in RBP binding depends on the precise nucleotide variant and its genomic context. Our results show that aggregating constraint at the level of transcripts is a well-powered intermediate approach to generate hypotheses for functional transcript binding using publicly available datasets.

Limitations of this study

Skipper analyzes read start signal in fixed windows using a simple discrete annotation framework. Other tools measure signal relative to features of interest such as exon-intron boundaries,[69,70] model signal with nucleotide level resolution,[9,10,58] reassign reads that map to multi-mapping genomic loci that do not appear to be expressed,[11] or weigh signal from other types of diagnostic events such as mismatches.[10] Users can blacklist reproducible regions of their choosing, but further work on best practices for blacklisting is warranted. Our *post hoc* fine-mapping broadens the utility of Skipper output but does not learn sequence motifs or binding affinity directly from CLIP data.[7,8] Furthermore, Skipper requires a matched input sample to construct the beta-binomial model for statistical testing: CLIP datasets lacking these controls currently cannot be analyzed using Skipper.

Our functional interrogation of translation factor binding sites is not definitive. The gapped k-mer SVM we use to model translation factor binding site occupancy is easy to use but less sophisticated than emerging tools based on deep learning.[53,58,71] We believe that future work will prove more sensitive to detect evolutionary constrained binding at the level of individual exons and transcripts. Our follow up work on RPS19 function relied on expressing variant open reading frames (ORFs) in HEK293T cells that may not reflect the true disease state, and eCLIP data alone do not answer whether observed changes in binding are due to differences in complex formation, localization, binding preference, or another mechanism altogether.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100317.

ACKNOWLEDGMENTS

This work was supported by grants R01 HG004659, R01 HG011864, RF1 MH126719, and U41/U24 HG009889 to G.W.Y. G.W.Y. is supported by an Allen Distinguished Investigator Award, a Paul G. Allen Frontiers Group advised grant of the Paul G. Allen Family Foundation.

E.A.B. is a Helen Hay Whitney Foundation Fellow. Data processing was performed on the Triton Shared Computing Cluster. This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from a National Institutes of Health SIG grant (#S10 OD026929). Data were also generated by the Sequencing Core Facility at the La Jolla Institute funded by National Institutes of Health SIG grant #S10 OD025052. Antibody illustration by Fredrik Edfors and light icon by laili at The Noun Project.

eQTL and sQTL data were generated by the Geno-type-Tissue Expression (GTEx) Project and retrieved from the GTEx Portal in January 2021. Diego Calderon, Ryan Marina, and Katherine Rothamel provided helpful comments on the Skipper pipeline, text, and figures. Steven Blue and Brian Yee assisted with eCLIP metadata curation.

AUTHOR CONTRIBUTIONS

E.A.B. conceived of the study, implemented the Skipper pipeline, performed RPS19 variant cloning, analyzed data, visualized results, and wrote the manuscript. H.H. compiled a draft of Skipper preprocessing steps and reviewed eCLIP data quality measures. J.R.M. generated ribosome protein eCLIP data and trained J.T.N and G.G.N. in the eCLIP technique. J.T.N. generated exogenously expressed RPS19 eCLIP data. G.G.N. generated EIF eCLIP data. G.W.Y. supervised the study and edited the manuscript.

DECLARATION OF INTERESTS

G.W.Y. is a cofounder, member of the board of directors, equity holder, and paid consultant for Locanabio and Eclipse BioInnovations and a distinguished visiting professor at the National University of Singapore. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict-of-interest policies.

Received: October 20, 2022
Revised: February 17, 2023
Accepted: April 6, 2023
Published: May 4, 2023

REFERENCES

1. Hentze, M.W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. Nat. Rev. Mol. Cell Biol. *19*, 327–341.
2. Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat. Rev. Genet. *15*, 829–845.
3. Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., and Zavolan, M. (2021). CLIP and complementary methods. Nature Reviews Methods Primers *1*, 1–23. https://doi.org/10.1038/s43586-021-00018-1.
4. Wheeler, E.C., Van Nostrand, E.L., and Yeo, G.W. (2018). Advances and challenges in the detection of transcriptome-wide protein–RNA interactions. Wiley Interdiscip. Rev. RNA *9*, e1436.
5. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods *13*, 508–514.
6. Uren, P.J., Bahrami-Samani, E., Burns, S.C., Qiao, M., Karginov, F.V., Hodges, E., Hannon, G.J., Sanford, J.R., Penalva, L.O.F., and Smith, A.D. (2012). Site identification in high-throughput RNA-protein interaction data. Bioinformatics *28*, 3013–3020. https://doi.org/10.1093/bioinformatics/bts569.
7. Katsantoni, M., van Nimwegen, E., and Zavolan, M. (2022). Improved analysis of (e)CLIP data with RCRUNCH yields a compendium of RNA-binding protein binding sites and motifs. Preprint at bioRxiv. https://doi.org/10.1101/2022.07.06.498949.
8. Feng, H., Bao, S., Rahman, M.A., Weyn-Vanhentenryck, S.M., Khan, A., Wong, J., Shah, A., Flynn, E.D., Krainer, A.R., and Zhang, C. (2019). Modeling RNA-binding protein specificity in vivo by precisely registering protein-RNA crosslink sites. Mol. Cell *74*, 1189–1204.e6. https://doi.org/10.1016/j.molcel.2019.02.002.
9. Krakau, S., Richard, H., and Marsico, A. (2017). PureCLIP: capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data. Genome Biol. *18*, 240. https://doi.org/10.1186/s13059-017-1364-2.
10. Drewe-Boss, P., Wessels, H.-H., and Ohler, U. (2018). omniCLIP: probabilistic identification of protein-RNA interactions from CLIP-seq data. Genome Biol. *19*, 183. https://doi.org/10.1186/s13059-018-1521-2.
11. Zhang, Z., and Xing, Y. (2017). CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. Nucleic Acids Res. *45*, 9260–9271. https://doi.org/10.1093/nar/gkx646.
12. Van Nostrand, E.L., Pratt, G.A., Yee, B.A., Wheeler, E.C., Blue, S.M., Mueller, J., Park, S.S., Garcia, K.E., Gelboin-Burkhart, C., Nguyen, T.B., et al. (2020). Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. Genome Biol. *21*, 90. https://doi.org/10.1186/s13059-020-01982-9.
13. Uhl, M., Tran, V.D., and Backofen, R. (2020). Improving CLIP-seq data analysis by incorporating transcript information. BMC Genom. *21*, 894. https://doi.org/10.1186/s12864-020-07297-0.
14. Wagner, S.D., Struck, A.J., Gupta, R., Farnsworth, D.R., Mahady, A.E., Eichinger, K., Thornton, C.A., Wang, E.T., and Berglund, J.A. (2016). Dose-dependent regulation of alternative splicing by MBNL proteins reveals biomarkers for myotonic dystrophy. PLoS Genet. *12*, e1006316. https://doi.org/10.1371/journal.pgen.1006316.

15. Becker, W.R., Jarmoskaite, I., Vaidyanathan, P.P., Greenleaf, W.J., and Herschlag, D. (2019). Demonstration of protein cooperativity mediated by RNA structure using the human protein PUM2. RNA *25*, 702–712.

16. Dassi, E. (2017). Handshakes and fights: the regulatory interplay of RNA-binding proteins. Front. Mol. Biosci. *4*, 67. https://doi.org/10.3389/fmolb.2017.00067.

17. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., et al. (2021). Sustainable data analysis with Snakemake. F1000Res. *10*, 33. https://doi.org/10.12688/f1000research.29032.2.

18. Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., et al. (2020). A large-scale binding and functional map of human RNA-binding proteins. Nature *583*, 711–719. https://doi.org/10.1038/s41586-020-2077-3.

19. Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Van Nostrand, E.L., Pratt, G.A., et al. (2018). Sequence, structure, and context preferences of human RNA binding proteins. Mol. Cell *70*, 854–867.e9. https://doi.org/10.1016/j.molcel.2018.05.001.

20. Jarmoskaite, I., Denny, S.K., Vaidyanathan, P.P., Becker, W.R., Andreasson, J.O.L., Layton, C.J., Kappel, K., Shivashankar, V., Sreenivasan, R., Das, R., et al. (2019). A quantitative and predictive model for RNA binding by human pumilio proteins. Mol. Cell *74*, 966–981.e18.

21. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773. https://doi.org/10.1093/nar/gky955.

22. Shah, A., Qian, Y., Weyn-Vanhentenryck, S.M., and Zhang, C. (2017). CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. Bioinformatics *33*, 566–567. https://doi.org/10.1093/bioinformatics/btw653.

23. Begg, B.E., Jens, M., Wang, P.Y., Minor, C.M., and Burge, C.B. (2020). Concentration-dependent splicing is enabled by Rbfox motifs of intermediate affinity. Nat. Struct. Mol. Biol. *27*, 901–912. https://doi.org/10.1038/s41594-020-0475-8.

24. Galarneau, A., and Richard, S. (2005). Target RNA motif and target mRNAs of the Quaking STAR protein. Nat. Struct. Mol. Biol. *12*, 691–698. https://doi.org/10.1038/nsmb963.

25. Zhang, Y., Cao, X., Gao, Z., Ma, X., Wang, Q., Cai, X., Zhang, Y., Zhang, Z., Wei, G., and Wen, B. (2022). MATR3-antisense LINE1 RNA meshwork scaffolds higher-order chromatin organization. Preprint at bioRxiv. https://doi.org/10.1101/2022.09.13.506124.

26. Xiong, F., Wang, R., Lee, J.-H., Li, S., Chen, S.-F., Liao, Z., Hasani, L.A., Nguyen, P.T., Zhu, X., Krakowiak, J., et al. (2021). RNA m6A modification orchestrates a LINE-1–host interaction that facilitates retrotransposition and contributes to long gene vulnerability. Cell Res. *31*, 861–885. https://doi.org/10.1038/s41422-021-00515-8.

27. Attig, J., Agostini, F., Gooding, C., Chakrabarti, A.M., Singh, A., Haberman, N., Zagalak, J.A., Emmett, W., Smith, C.W.J., Luscombe, N.M., and Ule, J. (2018). Heteromeric RNP assembly at LINEs controls lineage-specific RNA processing. Cell *174*, 1067–1081.e17. https://doi.org/10.1016/j.cell.2018.07.001.

28. Liu, N., Lee, C.H., Swigut, T., Grow, E., Gu, B., Bassik, M.C., and Wysocka, J. (2018). Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. Nature *553*, 228–232. https://doi.org/10.1038/nature25179.

29. Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N.M., and Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell *152*, 453–466.

30. Fasolo, F., Patrucco, L., Volpe, M., Bon, C., Peano, C., Mignone, F., Carninci, P., Persichetti, F., Santoro, C., Zucchelli, S., et al. (2019). The RNA-binding protein ILF3 binds to transposable element sequences in SINEUP lncRNAs. Faseb. J. *33*, 13572–13589. https://doi.org/10.1096/fj.201901618RR.

31. Thandapani, P., O'Connor, T.R., Bailey, T.L., and Richard, S. (2013). Defining the RGG/RG motif. Mol. Cell *50*, 613–623. https://doi.org/10.1016/j.molcel.2013.05.021.

32. Yagi, R., Miyazaki, T., and Oyoshi, T. (2018). G-quadruplex binding ability of TLS/FUS depends on the β-spiral structure of the RGG domain. Nucleic Acids Res. *46*, 5894–5901. https://doi.org/10.1093/nar/gky391.

33. Masuzawa, T., and Oyoshi, T. (2020). Roles of the RGG domain and RNA recognition motif of nucleolin in G-quadruplex stabilization. ACS Omega *5*, 5202–5208. https://doi.org/10.1021/acsomega.9b04221.

34. Lee, D.S.M., Ghanem, L.R., and Barash, Y. (2020). Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. Nat. Commun. *11*, 527. https://doi.org/10.1038/s41467-020-14404-y.

35. Ruggiero, E., Frasson, I., Tosoni, E., Scalabrin, M., Perrone, R., Marušič, M., Plavec, J., and Richter, S.N. (2022). Fused in liposarcoma protein, a new player in the regulation of HIV-1 transcription, binds to known and newly identified LTR G-quadruplexes. ACS Infect. Dis. *8*, 958–968. https://doi.org/10.1021/acsinfecdis.1c00508.

36. Butovskaya, E., Heddi, B., Bakalar, B., Richter, S.N., and Phan, A.T. (2018). Major G-quadruplex form of HIV-1 LTR reveals a (3 + 1) folding topology containing a stem-loop. J. Am. Chem. Soc. *140*, 13654–13662. https://doi.org/10.1021/jacs.8b05332.

37. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting splicing from primary sequence with deep learning. Cell *176*, 535–548.e24.

38. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F., and Guigó, R. (2021). Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. Nat. Commun. *12*, 727. https://doi.org/10.1038/s41467-020-20578-2.

39. Qi, T., Wu, Y., Fang, H., Zhang, F., Liu, S., Zeng, J., and Yang, J. (2022). Genetic control of RNA splicing and its distinct role in complex trait variation. Nat. Genet. *54*, 1355–1363. https://doi.org/10.1038/s41588-022-01154-4.

40. Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. *50*, 151–158.

41. Yang, E.-W., Bahn, J.H., Hsiao, E.Y.-H., Tan, B.X., Sun, Y., Fu, T., Zhou, B., Van Nostrand, E.L., Pratt, G.A., Freese, P., et al. (2019). Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. Nat. Commun. *10*, 1338.

42. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29. https://doi.org/10.1038/75556.

43. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

44. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425. https://doi.org/10.1016/j.cels.2015.12.004.

45. Adamson, S.I., Zhan, L., and Graveley, B.R. (2021). Functional characterization of splicing regulatory elements. Preprint at bioRxiv. https://doi.org/10.1101/2021.05.14.444228.

46. Rambout, X., Dequiedt, F., and Maquat, L.E. (2018). Beyond transcription: roles of transcription factors in pre-mRNA splicing. Chem. Rev. *118*, 4339–4364. https://doi.org/10.1021/acs.chemrev.7b00470.

47. Ma, J., Rong, L., Zhou, Y., Roy, B.B., Lu, J., Abrahamyan, L., Mouland, A.J., Pan, Q., and Liang, C. (2008). The requirement of the DEAD-box protein DDX24 for the packaging of human immunodeficiency virus type 1 RNA. Virology *375*, 253–264. https://doi.org/10.1016/j.virol.2008.01.025.

48. Zeng, M., Zhu, L., Li, L., and Kang, C. (2017). miR-378 suppresses the proliferation, migration and invasion of colon cancer cells by inhibiting SDAD1. Cell. Mol. Biol. Lett. *22*, 12. https://doi.org/10.1186/s11658-017-0041-5.

49. Thul, P.J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L.M., et al. (2017). A subcellular map of the human proteome. Science *356*, eaal3321. https://doi.org/10.1126/science.aal3321.

50. Samarsky, D.A., Fournier, M.J., Singer, R.H., and Bertrand, E. (1998). The snoRNA box C/D motif directs nucleolar targeting and also couples snoRNA synthesis and localization. EMBO J. *17*, 3747–3757. https://doi.org/10.1093/emboj/17.13.3747.

51. Young, D.J., Meydan, S., and Guydosh, N.R. (2021). 40S ribosome profiling reveals distinct roles for Tma20/Tma22 (MCT-1/DENR) and Tma64 (eIF2D) in 40S subunit recycling. Nat. Commun. *12*, 2976. https://doi.org/10.1038/s41467-021-23223-8.

52. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

53. Park, C.Y., Zhou, J., Wong, A.K., Chen, K.M., Theesfeld, C.L., Darnell, R.B., and Troyanskaya, O.G. (2021). Genome-wide landscape of RNA-binding protein target site dysregulation reveals a major impact on psychiatric disorder risk. Nat. Genet. *53*, 166–173. https://doi.org/10.1038/s41588-020-00761-3.

54. Lee, D., Gorkin, D.U., Baker, M., Strober, B.J., Asoni, A.L., McCallion, A.S., and Beer, M.A. (2015). A method to predict the impact of regulatory variants from DNA sequence. Nat. Genet. *47*, 955–961. https://doi.org/10.1038/ng.3331.

55. Ulirsch, J.C., Verboon, J.M., Kazerounian, S., Guo, M.H., Yuan, D., Ludwig, L.S., Handsaker, R.E., Abdulhay, N.J., Fiorini, C., Genovese, G., et al. (2019). The genetic landscape of diamond-blackfan anemia. Am. J. Hum. Genet. *104*, 356. https://doi.org/10.1016/j.ajhg.2018.12.011.

56. Da Costa, L., Tchernia, G., Gascard, P., Lo, A., Meerpohl, J., Niemeyer, C., Chasis, J.-A., Fixler, J., and Mohandas, N. (2003). Nucleolar localization of RPS19 protein in normal cells and mislocalization due to mutations in the nucleolar localization signals in 2 Diamond-Blackfan anemia patients: potential insights into pathophysiology. Blood *101*, 5039–5045. https://doi.org/10.1182/blood-2002-12-3878.

57. Devlin, E.E., Dacosta, L., Mohandas, N., Elliott, G., and Bodine, D.M. (2010). A transgenic mouse model demonstrates a dominant negative effect of a point mutation in the RPS19 gene associated with Diamond-Blackfan anemia. Blood *116*, 2826–2835. https://doi.org/10.1182/blood-2010-03-275776.

58. Ghanbari, M., and Ohler, U. (2020). Deep neural networks for interpreting RNA-binding protein target preferences. Genome Res. *30*, 214–226. https://doi.org/10.1101/gr.247494.118.

59. Denichenko, P., Mogilevsky, M., Cléry, A., Welte, T., Biran, J., Shimshon, O., Barnabas, G.D., Danan-Gotthold, M., Kumar, S., Yavin, E., et al. (2019). Specific inhibition of splicing factor activity by decoy RNA oligonucleotides. Nat. Commun. *10*, 1590.

60. Arandel, L., Matloka, M., Klein, A.F., Rau, F., Sureau, A., Ney, M., Cordier, A., Kondili, M., Polay-Espinoza, M., Naouar, N., et al. (2022). Reversal of RNA toxicity in myotonic dystrophy via a decoy RNA-binding protein with high affinity for expanded CUG repeats. Nat. Biomed. Eng. *6*, 207–220. https://doi.org/10.1038/s41551-021-00838-2.

61. Jackson, A.L., Burchard, J., Schelter, J., Chau, B.N., Cleary, M., Lim, L., and Linsley, P.S. (2006). Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. RNA *12*, 1179–1187. https://doi.org/10.1261/rna.25706.

62. Zhang, X., Wang, W., Zhu, W., Dong, J., Cheng, Y., Yin, Z., and Shen, F. (2019). Mechanisms and functions of long non-coding RNAs at multiple regulatory levels. Int. J. Mol. Sci. *20*, 5573. https://doi.org/10.3390/ijms20225573.

63. Rom, A., Melamed, L., Gil, N., Goldrich, M.J., Kadir, R., Golan, M., Biton, I., Perry, R.B.-T., and Ulitsky, I. (2019). Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. Nat. Commun. *10*, 5092. https://doi.org/10.1038/s41467-019-13075-8.

64. Ludwig, L.S., Gazda, H.T., Eng, J.C., Eichhorn, S.W., Thiru, P., Ghazvinian, R., George, T.I., Gotlib, J.R., Beggs, A.H., Sieff, C.A., et al. (2014). Altered translation of GATA1 in Diamond-Blackfan anemia. Nat. Med. *20*, 748–753. https://doi.org/10.1038/nm.3557.

65. Da Costa, L., Leblanc, T., and Mohandas, N. (2020). Diamond-Blackfan anemia. Blood *136*, 1262–1273. https://doi.org/10.1182/blood.2019000947.

66. Moras, M., Lefevre, S.D., and Ostuni, M.A. (2017). From erythroblasts to mature red blood cells: organelle clearance in mammals. Front. Physiol. *8*, 1076. https://doi.org/10.3389/fphys.2017.01076.

67. Mortensen, M., Ferguson, D.J.P., Edelmann, M., Kessler, B., Morten, K.J., Komatsu, M., and Simon, A.K. (2010). Loss of autophagy in erythroid cells leads to defective removal of mitochondria and severe anemia in vivo. Proc. Natl. Acad. Sci. USA *107*, 832–837. https://doi.org/10.1073/pnas.0913170107.

68. Doulatov, S., Vo, L.T., Macari, E.R., Wahlster, L., Kinney, M.A., Taylor, A.M., Barragan, J., Gupta, M., McGrath, K., Lee, H.-Y., et al. (2017). Drug discovery for Diamond-Blackfan anemia using reprogrammed hematopoietic progenitors. Sci. Transl. Med. *9*, eaah5645. https://doi.org/10.1126/scitranslmed.aah5645.

69. Her, H.-L., Boyle, E., and Yeo, G.W. (2022). Metadensity: a background-aware python pipeline for summarizing CLIP signals on various transcriptomic sites. Bioinform. Adv. *2*, vbac083. https://doi.org/10.1093/bioadv/vbac083.

70. Yee, B.A., Pratt, G.A., Graveley, B.R., Van Nostrand, E.L., and Yeo, G.W. (2019). RBP-Maps enables robust generation of splicing regulatory maps. RNA *25*, 193–204. https://doi.org/10.1261/rna.069237.118.

71. Horlacher, M., Wagner, N., Moyon, L., Kuret, K., Goedert, N., Salvatore, M., Ule, J., Gagneur, J., Winther, O., and Marsico, A. (2022). Towards In-Silico CLIP-Seq: Predicting Protein-RNA Interaction via Sequence-To-Signal Learning. Preprint at bioRxiv. https://doi.org/10.1101/2022.09.16.508290.

72. Lovci, M.T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T.Y., Stark, T.J., Gehman, L.T., Hoon, S., et al. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat. Struct. Mol. Biol. *20*, 1434–1442. https://doi.org/10.1038/nsmb.2699.

73. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). Genome project data processing subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

74. Jiang, H., Lei, R., Ding, S.-W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinf. *15*, 182. https://doi.org/10.1186/1471-2105-15-182.

75. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

76. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

77. Liu, D. (2019). Algorithms for efficiently collapsing reads with unique molecular identifiers. PeerJ *7*, e8275. https://doi.org/10.7717/peerj.8275.

78. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast

universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

79. Lee, D. (2016). LS-GKM: a new gkm-SVM for large-scale datasets. Bioinformatics *32*, 2196–2198. https://doi.org/10.1093/bioinformatics/btw142.
80. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419. https://doi.org/10.1038/nmeth.4197.
81. Yee, T.W. (2015). Vector Generalized Linear and Additive Models (New York: Springer). https://doi.org/10.1007/978-1-4939-2818-7.
82. Krijthe. (2015). Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. R package version 0.13. https://github. com/jkrijthe.
83. Rasheedi, S., Shun, M.C., Serrao, E., Sowd, G.A., Qian, J., Hao, C., Dasgupta, T., Engelman, A.N., and Skowronski, J. (2016). The cleavage and polyadenylation specificity factor 6 (CPSF6) subunit of the capsid-recruited pre-messenger RNA cleavage factor I (CFIm) complex mediates HIV-1 integration into genes. J. Biol. Chem. *291*, 11809–11819.
84. Aznarez, I., Barash, Y., Shai, O., He, D., Zielenski, J., Tsui, L.C., Parkinson, J., Frey, B.J., Rommens, J.M., and Blencowe, B.J. (2008). A systematic analysis of intronic sequences downstream of 5′ splice sites reveals a widespread role for U-rich motifs and TIA1/TIAL1 proteins in alternative splicing regulation. Genome Res. *18*, 1247–1258.
85. Blue, S.M., Yee, B.A., Pratt, G.A., Mueller, J.R., Park, S.S., Shishkin, A.A., Starner, A.C., Van Nostrand, E.L., and Yeo, G.W. (2022). Transcriptome-wide identification of RNA-binding protein binding sites using seCLIP-seq. Nat. Protoc. *17*, 1223–1265. https://doi.org/10.1038/s41596-022-00680-z.
86. Anger, A.M., Armache, J.-P., Berninghausen, O., Habeck, M., Subklewe, M., Wilson, D.N., and Beckmann, R. (2013). Structures of the human and Drosophila 80S ribosome. Nature *497*, 80–85. https://doi.org/10.1038/nature12104.
87. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940–3941. https://doi.org/10.1093/bioinformatics/bti623.
88. Pronobis, M.I., Deuitch, N., and Peifer, M. (2016). The Miraprep: a protocol that uses a miniprep Kit and provides maxiprep yields. PLoS One *11*, e0160509. https://doi.org/10.1371/journal.pone.0160509.
89. Li, L., Huang, K.-L., Gao, Y., Cui, Y., Wang, G., Elrod, N.D., Li, Y., Chen, Y.E., Ji, P., Peng, F., et al. (2021). An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. Nat. Genet. *53*, 994–1005. https://doi.org/10.1038/s41588-021-00864-5.
90. Mittleman, B.E., Pott, S., Warland, S., Zeng, T., Mu, Z., Kaur, M., Gilad, Y., and Li, Y. (2020). Alternative polyadenylation mediates genetic regulation of gene expression. Elife *9*, e57492. https://doi.org/10.7554/eLife.57492.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit anti-eIF2B5	Bethyl Laboratories	Cat# A302-556A; RRID:AB_2034833
Rabbit anti-Ligatin	Bethyl Laboratories	Cat# A303-006A; RRID:AB_10750476
Rabbit anti-eIF2beta/EIF2S2	Bethyl Laboratories	Cat# A301-743A; RRID:AB_1210964
Rabbit anti-eIF3J/EIF3S1	Bethyl Laboratories	Cat# A301-746A; RRID:AB_1210975
Rabbit anti-RPS14	Bethyl Laboratories	Cat# A304-031A; RRID:AB_2621280
Rabbit anti-RPL35A/Ribosomal Protein L35a	Bethyl Laboratories	Cat# A305-106A; RRID:AB_2631501
Rabbit anti-Ribosomal Protein S3A/RPS3A	Bethyl Laboratories	Cat# A305-001A; RRID:AB_2621195
Rabbit anti-RPL29/ Ribosomal Protein L29	Bethyl Laboratories	Cat# A305-056A; RRID:AB_2621250
Rabbit anti-RPS19	Bethyl Laboratories	Cat# A304-002A; RRID:AB_2620351
Rabbit anti-V5 Tag	Bethyl Laboratories	Cat# A190-120A; RRID:AB_67586
Critical commercial assays		
Q5® Site-Directed Mutagenesis Kit	NEB	E0554S
Deposited data		
Translation factor eCLIP	This paper	GEO: GSE213867
ENCODE 3 CLIP Skipper code and output	This paper	Figshare: https://figshare.com/articles/dataset/Skipper_RNA-protein_interaction_profiles/21206009 (https://doi.org/10.6084/m9.figshare.21206009.v1 and https://doi.org/10.6084/m9.figshare.21272991.v1)
RPS19-V5 eCLIP	This paper	GEO: GSE224998 Figshare: https://figshare.com/articles/dataset/RPS19_construct_eCLIP_data/22097072
Experimental models: Cell lines		
Human Lenti-X™ 293T Cell Line	Takara Bio USA	632180
Human K562	ATCC	Related to CCL-243
Oligonucleotides		
RPS19_SDM_R62W_ F: TTCCACAGCGtGGCACCTGTA		N/A
RPS19_SDM_R62W_R: GCAGCTCGCGTGTAGAAC		N/A
RPS19_SDM_R101H_F: AGTGTGGCCCaCCGGGTCCTC		N/A
RPS19_SDM_R101H_R: CTTGGAGCCTCGGCTGAAG		N/A
Recombinant DNA		
p223-RPS19	Orfeome v8.1	BC000023
p223-RPS19-R62W	This paper	N/A
p223-RPS19-R101H	This paper	N/A
Software and algorithms		
Skipper v1.0.0	This paper	https://github.com/YeoLab/skipper/
Piranha v1.2.1	Uren et al.[6]	http://smithlabresearch.org/software/piranha/
CLIPper v1.0 (and merge_peaks)	Lovci et al.[72]	https://github.com/YeoLab/clipper; https://github.com/YeoLab/merge_peaks
CTK v1.1.4	Shah et al.[22]	https://zhanglab.c2b2.columbia.edu/index.php/ECLIP_data_analysis_using_CTK
omniCLIP v0.2.0	Drewe-Boss et al.[10]	https://github.com/philippdre/omniCLIP
PureCLIP v1.3.1	Krakau et al.[9]	https://github.com/skrakau/PureCLIP
samtools v1.15.1	Li et al.[73]	http://www.htslib.org/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Skewer v0.2.2	Jiang et al.[74]	https://github.com/relipmoc/skewer
Fastp v0.23.2	Chen et al.[75]	https://github.com/OpenGene/fastp
bedtools v2.30.0	Quinlan et al.[76]	https://github.com/arq5x/bedtools2
UMICollapse	Liu, 2019[77]	https://github.com/Daniel-Liu-c0deb0t/UMICollapse
STAR v2.7.10a	Dobin et al.[78]	https://github.com/alexdobin/STAR
LS-GKM v0.1.1	Lee, 2016[79]	https://github.com/Dongwon-Lee/lsgkm

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. Gene Yeo (geneyeo@ucsd.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- ENCODE eCLIP fastqs and CLIPper processed data are available on the ENCODE Project website: https://www.encodeproject.org/encore-matrix/?type=Experiment&status=released&internal_tags=ENCORE
- Translation eCLIP raw fastqs and reproducible enriched window output are available at GEO: GSE213867 and GSE224998.
- Additional summary data for Skipper output are available on Figshare (https://doi.org/10.6084/m9.figshare.21206009.v1): https://figshare.com/articles/dataset/Skipper_RNA-protein_interaction_profiles/21206009
- The Skipper pipeline including example input are available under a BSD license at https://github.com/YeoLab/skipper/ and deposited on Figshare.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

K562 (chronic myelogenous leukemia; 53-year-old female) was cultured in RPMI1640 with 10% FBS at 37C. HepG2 (hepatocellular carcinoma; 15-year-old male) was cultured in DMEM with 10% FBS at 37C. K562 was authenticated by STR profiling through ATCC. HepG2 was not authenticated.

METHOD DETAILS

Tiling windows across the transcriptome

HepG2 and K562 total RNA-seq were downloaded from the encodeproject.org website. Transcript abundance was evaluated using Salmon[80] per documented guidelines. GENCODE version 38 gene annotations were downloaded from the gencodegenes.org website, and transcripts with less than 1 transcript per million were filtered using the pyranges package in our custom script subset_gff.py for K562 and HepG2 cell lines separately.

We use the custom script parse_gff.R with a manual rankings of GENCODE accession types (roughly small noncoding RNAs first, then mRNAs, then lncRNAs, then pseudogenes) and feature types (ranges containing small noncoding RNA exons first, then both start and stop codons, start codons, stop codons, other CDS regions, 3′ UTRs, 5′ UTRs, noncoding isoforms of mRNA, lncRNA exons, 3′ and 5′ splice sites within 100 nucleotides, 3′ splice sites only, 5′ splice sites only, 3′ and 5′ splice sites within 500 nucleotides, 3′ splice sites only, 5′ splice sites only, primary transcript miRNAs, and finally introns) to tile windows. We iterate over ranked features and retrieve contiguous coordinates that are split into evenly sized windows not exceeding 100 nucleotides in length. For example, a 50-nucleotide exon would yield a 50-nucleotide long window, but a 210-nucleotide exon would be split into three 70-nucleotide windows. The resulting disjoint windows are annotated with all overlapping transcripts and features and numbered uniquely.

Skipper data processing

Skipper utilizes a single manifest of samples to preprocess, model, visualize, and summarize read start signal from any CLIP-seq data with matched input samples. For preprocessing, Skipper trims fastq files using skewer,[74] cuts and pastes unimolecular identifiers using fastp,[75] aligns reads with STAR,[78] and deduplicates unimolecular identifiers using UMIcollapse.[77]

For beta-binomial modeling, counts of 5′ ends of reads per strand and GC content per tiled window and RepeatMasker element are calculated using SAMtools[73] and bedtools.[76] Genomic positions that span multiple repetitive elements are ignored. Read counts are placed into ten GC content bins for genome-mapped reads and twenty bins for repetitive elements where elements are binned in accordance with the average GC content of all instances of a particular repetitive element. Enrichment odds ratios are calculated by adding a pseudocount equal to the average rate of success per bin to avoid division by zero. A null overdispersion parameter (*rho*) and mean fold change parameter (*mu*) is estimated across bins by $\langle r_1, r_2 \rangle \sim s_{1,b}/s_{2,b}$ via the vglm function and beta-binomial family from the VGAM package,[81] where r_i is the counts per window or element in replicate *i*, and $s_{i,b}$ are the sums of counts in replicate *i* in GC bin *b*. With the overdispersion parameter estimated under the null, p values are calculated for immunoprecipitated versus input samples with the *pbetabinom* function from VGAM. P-values less than 10^{-12} were replaced with 10^{-12} to address floating point imprecision. A false discovery rate is enforced by filtering windows for the sum of immunoprecipitated and input reads passing a dynamically determined threshold to maximize the number of hits and using the *p.adjust* function in R. Windows and repetitive elements that passed a 20% FDR in both replicates are called as reproducibly enriched, but reproducible enriched windows that were called in more than 17% of either HepG2 or K562 eCLIP samples are blacklisted and removed. Concordance between pairs of replicates is then assessed by Fisher's Exact Test.

Automated analysis of Skipper's reproducible enriched windows

For clustering, transcriptomic windows counts are summarized as belonging to one of the following ranked categories: rRNA, snoRNA, snRNA, MtRNA, 7SK, Y RNA, and finally a combination of transcript type defined by the GENCODE gff and the feature type as defined above. Because of the large number of simple repeats, only repetitive elements enriched by at least 2.5 units of $\log_2$ fold change are included in clustering. Repeats that are not one of snRNAs, Y RNAs, tRNAs, LINE 1 sense or antisense, Alu sense or antisense, LTR, 7SK, or another LINE element are grouped together as ''Other repetitive elements''. Each category is assigned an entropy contribution defined at $p_i \log_2 \frac{p_i}{q_i}$ where p is the fraction of windows for category *i* in the queried CLIP sample and q is the fraction of windows in category *i* across all CLIPs. Reference data was clustered using Pearson correlation distance and the McQuitty agglomeration method via the *hclust* function in R. The nine classes of RBPs were created by cutting the tree into ten subgroups and reassigning a singleton clade. Each class was labeled according to the category with the greatest entropy contribution.

Skipper reproducible enriched windows are fine mapped by recentering on local maxima of binding enrichment (i.e. immunoprecipitated over input reads summed over 75 nt intervals) and extended to a fixed 75 nt size. 75 nt windows overlapping the original reproducible enriched windows are iteratively selected until no local maxima remain or enrichment falls below the median across all positions. For motif calling, 75 nt control windows are created first by selecting a random window in the partitioned transcriptome of the same feature group – noncoding exons, mRNA exons, proximal introns (within 500 nt of exon-intron boundary), or distal introns – and then randomly selecting a center. The findMotifsGenome.pl script from HOMER is run on fine-mapped windows with the following options: -preparsedDir <directory> -size given -rna -nofacts -S 20 -len 5,6,7,8,9 -nlen 1 -bg <background windows>

For gene ontology analysis, windows lying in genes belonging to each term are tallied and compared to the representation of each term across all ENCODE 3 CLIPs to control for expression, library preparation bias, and mappability. P values are calculated by binomial test with the rate of success equal to the representation of the term in the ENCODE 3 CLIPs. P-values are Bonferroni corrected for multiple hypotheses. For t-SNE visualization, tallies of feature type and transcript type windows from CLIP-seq query data are pooled with ENCODE reference data and plotted in two dimensions using the Rtsne package.[82]

Use of CLIP analysis tools

To prioritize transcript annotations for all tools, the gene_type and transcript_type fields in GENCODE were manually ranked and stored in the accession_type_rankings.txt file. Piranha v1.2.1 (no covariates) and CTK v1.1.4 (eCLIP with statistical significance) were downloaded and run by following instructions on their corresponding lab software webpages. Counts were aggregated using Skipper tiled windows for Piranha, as the instructions did not recommend a procedure for calling peaks or aggregating counts. PureCLIP v1.3.1 was installed using Conda with an activated Bioconda channel as directed on its GitHub page. We followed the PureCLIP Read the Docs page for incorporating input control data, adding the -nt 12 option to enable completion within 2 hours per CLIP. We downloaded omniCLIP v0.2.0 via GitHub and edited the source code as suggested by users on the GitHub Issues page to allow it to compile. generateDB, parsingBG, and parsingCLIP were run as directed on the GitHub. The run_omniCLIP command was used with the –nb-cores 12 option to enable completion within 12 hours per CLIP. CLIPper v1.0-processed IDR peaks were downloaded from the ENCODE Project website.

For CTK, CLIPper, and omniCLIP, the center of each hit region was overlapped with the Skipper transcriptome-tiled windows For PureCLIP, the called crosslink site was used.

For CLIPper, CTK, PureCLIP, windows containing hits from both replicates were called candidate binding sites. omniCLIP merges replicates and outputs one file of hits which was used for candidate binding sites.

Precision and relative recall calculations for CLIP-seq benchmarking

FASTKD2 enriched windows were ascertained as true positive binding sites if they aligned to chrM and false positives otherwise.[18] Scripts for assessing PUM2 binding affinity[20] were modified to score arbitrary sequences. Enriched windows and matched control regions were extended 7 nt and scored. Scores above the 95th percentile in the control regions were ascertained as true positives and

scores below the 50th percentile as true negatives. PRPF8 enriched windows were ascertained as true positives if they were annotated as GENCODE 5′ splice site proximal (within 500 nt) and true negatives otherwise. TARDBP enriched windows were extended 5 nt and queried using bedtools nuc for the GURUG motif. Enriched windows containing the GURUG motif were ascertained as true positives and true negatives otherwise.[18] TROVE2 enriched windows were ascertained as true positives if they derived from Y RNAs and true negatives otherwise.[12]

CLIPper peaks are much narrower than the tiled transcriptomic windows used by Skipper and do not respect boundaries around UTRs and exon junctions. CLIPper peaks were reassigned the transcriptomic window that overlapped the center of the CLIPper peak. Thus, multiple CLIPper peaks within the same transcriptomic window were not double counted.

Because the full set of all true binding events is not known, we calculated a relative recall measure for each CLIP analysis method as follows: the number of ascertained true positives detected by the method divided by the number of unique true positives ascertained across all methods.

RBFOX2 knockdown-sensitive exon analysis

RBFOX2 knockdown-sensitive exons were downloaded from Van Nostrand et al. 2016 by downloading source data to Figure S10 (mislabeled as source data for Figure S13 on the Nature Methods online article) and lifted over to GRCh38. Skipper windows or CLIPper peaks within 500 nucleotides of knockdown-sensitive exons were retrieved using bedtools flank and intersect commands and labeled with the corresponding exon SepScore.

LINE1 evolutionary analysis

LINE1 specificity to primates was defined as the percent of individual instances of each LINE1 type that were novel to primates.[27] The GC-corrected $\log_2$ enrichment per element reported by Skipper was then plotted against the specificity to primates.

Evaluating RBP binding near alternative splice sites

Total RNA-seq for HepG2 and K562 cell lines was downloaded from the ENCODE Project website encodeproject.org. Alternative splicing was assessed using rMATS with the following options: –gtf gencode.v38.annotation.gtf –bi <STAR reference> -t paired –readLength 50 –od rmats –statoff. Alternative 5′ and -3′ splice sites were converted to BED format and intersected with all reproducible enriched windows called using ENCODE Project eCLIP data. Alternative exons with fewer than 20 total exon junction reads were discarded. Bias toward alternative or constitutive splice sites was calculated for each eCLIP experiment using the *binom.test* function with a probability of success (enriched windows overlapping the alternative splice site) of 0.5. Testing alternative versus constitutive splice site binding bias stratified by alternative exon usage was performed using the *chisq.test* function.

Known regulators of RNA splicing and decay were identified as belonging to the Gene Ontology terms "RNA splicing" and "Post-transcriptional regulation of gene expression", plus CPSF6. TIAL1 and CPSF6 were not annotated with RNA splicing or post-transcriptional regulation of gene expression, but are well known to play those respective roles.[83,84]

Assessing subcellular localization

RBP subcellular localization[49] was downloaded from the Human Protein Atlas website proteinatlas.org. RBPs were noted for binding to the nucleoli, nucleoplasm, cytosol, or mitochondria in either the "Main location" or "Additional location" fields.

eCLIP of translation factors

eCLIP was performed according to our published protocol.[85] eCLIPs were performed using antibodies from Bethyl Laboratories against EIF2B5 (A302-556A), EIF2D (A303-006A), EIF2S2 (A301-743A), EIF3J (A301-746A), RPS14 (A304-031A), RPL35A (A305-106A), RPS3A (A305-001A), RPL29 (A305-056A), and RPS19 (A304-002A) in K562 cells. The ribosome schematic was derived from entry 4V6X on PDB.[86]

Selective constraint testing

LS-GKM[79] was trained on 75-nt fine-mapped windows centered on binding signal within Skipper reproducibly enriched windows for translation factor eCLIPs and approximately 20,000 randomly sampled control windows used as background. Gapped kmer SVMs were trained using the following command: *gkmtrain -m 6000 -l 10 -t 3 -k 6 ${id}.finemapped_windows.fa ${id}.sampled_windows.fa gkm_models/${id}*. Area under the precision recall curve was assessed via the ROCR R package[87] using 5-fold cross validation estimates.

Genetic variants in gnomAD overlapping 75-nt fine-mapped windows were queried remotely using bcftools looping across CLIP experiments and chromosomes:

bcftools query -R $CLIP_bed -f '%CHROM\t%POS\t%ID\t%REF\t%ALT\t%INFO/AC\t%INFO/AN\n' $https://gnomad-public-us-east-1.s3.amazonaws.com/release/3.1.2/vcf/genomes/gnomad.genomes.v3.1.2.sites.${chromosome}.vcf.bgz

Variant fasta files were created from reference fastas by substituting single nucleotide variants one at a time. Variant and reference fastas were scored for all chromosomes using the gkmpredict command: *gkmpredict -T 1 gnomad/${id}/${chromosome}.mut.fa gkm_models/${id}.model.txt gkm_out/${id}.${chromosome}.mut.txt*. Delta scores were computed as the difference between gkm-SVM predictions on variant and reference fastas. Variants were placed into four bins: singletons, allele frequency < 0.1%, allele

frequency between 0.1% and 1%, and allele frequency >1%, and delta scores were regressed against bin rank using the *lm* command in R to determine relative constraint. To assess significance, delta scores were permuted within allele frequency bins 5000 times to create a null distribution of relative constraint such that constrained transcripts were required to exceed the transcriptome-wide average trend. Instances in which no permutation exceeded the relative constraint of the observed values were replaced with 10^{-4}, and the *p.adjust* function in R was used to enforce a 10% false discovery rate.

Enrichment for feature types in constrained transcripts was detected by Fisher's Exact Test stratifying enriched window counts by whether the transcript passed statistical significance and whether the window derived from a particular feature type.

Wild type and mutant RPS19 plasmid cloning

The RPS19 ORF from the Orfeome 8.1 collection was inoculated into LB with 200 ug/mL hygromycin b and purified by Qiagen miniprep. We cloned wild type RPS19 into pEF5/FRT/V5-DEST in a single step LR Clonase reaction with 100 ng RPS19 ORF plasmid, 300 ng pDONR 221 Vector, 300 ng pEF5/FRT/V5-DEST, and 2 uL of LR Clonase II enzyme mix (Thermo 11791020) in a 10 uL reaction volume and incubated for 2 hours at room temperature before adding 1 uL proteinase K solution and incubating at 37C for ten minutes. Gateway product was transformed into One Shot Stbl3 chemically competent cells (Thermo C737303) by heat shock for 45 seconds at 42C. Cultures were purified by Miraprep.[88] Site-directed mutagenesis primers were designed using NEBaseChanger to create R62W and R101H constructs. Mutations were induced using the Q5 Site-Directed Mutagenesis Kit (NEB E0554S) according to manufacturer's instructions. All plasmids were sequenced by Primordium Labs.

eCLIP of wild type and mutant V5-tagged RPS19

Twelve 10 cm plates were seeded with Lenti-X^{TM} 293T cells in 10 mL of DMEM media. Plated cells grew undisturbed until ~80% confluency as determined by microscopy. Then cells were transfected with the Lipofectamine^{TM} 3000 Transfection Reagent (Invitrogen); first, 30 uL of each V5 tagged RPS19 plasmid (WT, R62W, R101H) was added to 84 uL of p3000 reagent followed by 130.2 uL of Lipofectamine 3000 under a laminar flow hood and vortexed for 15 seconds. Tubes then incubated at room temperature for 15 minutes. 80 uL of each plasmid transfection mix was added dropwise to the plated cells by micropipette. Cells for eCLIP size-matched inputs were not transfected. 48 hours after the transfection, cells were irradiated at 400 mJ/Cm^2 in a UV crosslinker and pelleted. Cell pellets were stored at -80C before proceeding with the eCLIP protocol with anti-V5 antibody (Bethyl Laboratories A190-120A).

RPS19 construct eCLIP analysis

Skipper was run by pairing each CLIP sample with one of the three input samples and estimating overdispersion from the three input samples. tRNA enrichment was evaluated per replicate by performing a paired t-test on mutant enrichments versus wild type enrichments for all tRNA genes. Gene-wise signal was aggregated by taking the sum of the $\log_2$ enrichments for all enriched windows per gene separately for intronic windows and CDS windows. Correlation was calculated by using the *cor* function in R with the complete=TRUE option.

Overlaps between enriched windows and QTLs

eQTL and sQTL[40] v8 datasets were downloaded from the GTEx portal. SNPs with the most significant p-value per gene or splice graph were retained as lead eQTL or sQTL, respectively. Lead 3′ alternate polyadenylation QTL (3aQTL)[89] were downloaded from Synapse ID syn22131046. Polyadenylation sites[90] were downloaded from GEO ID GSE138197. Skipper enriched windows and windows containing CLIPper IDR peaks were intersected SNP positions using bedtools.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests and test details are listed in the results section. Significance testing for beta-binomial distributions was performed using the VGAM package in R. Other statistical tests were performed using base R functions. See method details for more information.